

# ТРАЕКТОРИЯ ИССЛЕДОВАНИЙ — ЧЕЛОВЕК, ПРИРОДА, ТЕХНОЛОГИИ



НАУЧНЫЙ ПЕРИОДИЧЕСКИЙ ЖУРНАЛ

№ 3 (3) 2022







# ТРАЕКТОРИЯ ИССЛЕДОВАНИЙ – ЧЕЛОВЕК, ПРИРОДА, ТЕХНОЛОГИИ

№ 3 (3) 2022

## Главный редактор

Ярмошенко Илья Владимирович, к. ф.-м. н., директор Института промышленной экологии УрО РАН (г. Екатеринбург)

## Заместитель главного редактора

Васильев Алексей Владимирович, к. т. н., заведующий лабораторией Института промышленной экологии УрО РАН (г. Екатеринбург)

## Редакционная коллегия

Вараксин Анатолий Николаевич, д. ф.-м. н., профессор, главный научный сотрудник Института промышленной экологии УрО РАН (г. Екатеринбург)

Веселкин Денис Васильевич, д. б. н., заведующий лабораторией Института экологии растений и животных УрО РАН (г. Екатеринбург)

Екидин Алексей Акимович, к. ф.-м. н., доцент, ведущий научный сотрудник Института промышленной экологии УрО РАН (г. Екатеринбург)

Захаров Вячеслав Иосифович, д. ф.-м. н., профессор, заведующий лабораторией Уральского федерального университета им. первого Президента России Б. Н. Ельцина (г. Екатеринбург)

Иванов Владимир Юрьевич, к. ф.-м. н., директор института Уральского федерального университета им. первого Президента России Б. Н. Ельцина (г. Екатеринбург)

Крышев Александр Иванович, д. б. н., заведующий лабораторией Научно-производственного объединения «Тайфун» Росгидромета (г. Обнинск)

Марков Дмитрий Владимирович, д. т. н., директор АО «ГНЦ РФ ТРИНИТИ» (г. Троицк)

Микляев Петр Сергеевич, д. г.-м. н., профессор РАН, главный научный сотрудник Института геоэкологии им. Е. М. Сергеева РАН (г. Москва)

Сюрдо Александр Иванович, д. ф.-м. н., профессор, главный научный сотрудник Института физики металлов им. М. Н. Михеева УрО РАН (г. Екатеринбург)

Талалай Александр Григорьевич, д. г.-м. н., профессор, заведующий кафедрой Уральского государственного горного университета (г. Екатеринбург)

Тузанкина Ирина Александровна, заслуженный деятель науки РФ, д. м. н., профессор, Институт иммунологии и физиологии УрО РАН (г. Екатеринбург)

Юрков Анатолий Константинович, к. г.-м. н., заведующий лабораторией Института геофизики им. Ю. П. Булашевича УрО РАН (г. Екатеринбург)



## СОДЕРЖАНИЕ

**ТРАЕКТОРИЯ  
ИССЛЕДОВАНИЙ –  
ЧЕЛОВЕК, ПРИРОДА,  
ТЕХНОЛОГИИ**

**№ 3 (3) 2022**

От выпускающих редакторов <b>Е. Д. Константинова, А. П. Сергеев</b> .....	3
Методы визуализации данных в медико-биологических исследованиях <b>С. Ю. Огородникова, Е. Д. Константинова</b> .....	4
Линейная и нелинейная регрессия в биологии и медицине: подход, основанный на скользящем среднем <b>А. Н. Вараксин, Ю. В. Шалаумова, Т. А. Маслакова</b> .....	19
Разработка метода управления активным экзоскелетом для реабилитации моторной деятельности верхних конечностей человека <b>В. М. Антипов, А. А. Бадарин, А. В. Андреев, В. В. Грубов, Е. Н. Пицик, С. А. Лобов, В. А. Максименко, В. Б. Казанцев, А. Е. Храмов</b> .....	39
Определение момента времени начала совершения воображаемого движения для задач реабилитации <b>В. С. Хорев, Н. В. Смирнов, С. А. Куркин, А. П. Сергеев, С. Ю. Гордлева, А. Е. Храмов</b> .....	48
Вертикальное распределение характеристик запыленности приземного слоя атмосферного воздуха г. Екатеринбурга <b>И. Е. Субботина, М. С. Ремезова, А. Г. Бувевич, А. П. Сергеев, Е. М. Баглаева, А. В. Шичкин, А. С. Буторова, М. В. Сергеева</b> .....	60
Сравнительный анализ алгоритмов разделения данных для обучения искусственных нейронных сетей при моделировании пространственного распределения тяжелых металлов в верхнем слое почвы <b>Е. М. Баглаева, А. П. Сергеев, А. В. Шичкин, А. Г. Бувевич, И. Е. Субботина, А. С. Буторова</b> .....	73
Совместное действие бинарных факторов и равенство рисков <b>Ю. В. Нагребцкая, А. И. Фаткуллина</b> .....	89
Анализ дискретных данных. Основные понятия <b>В. Г. Панов</b> .....	105
Предисловие к статье В. Н. Чуканова «Экологические проблемы – системный подход» <b>А. Н. Вараксин</b> .....	142
Экологические проблемы – системный подход <b>В. Н. Чуканов</b> .....	143

**Главный редактор:**  
И. В. Ярмошенко

**Редактор:**  
М. Э. Чупрякова

**Выпускающие редакторы:**  
Е. Д. Константинова  
А. П. Сергеев

**Макетирование  
и компьютерная верстка:**  
М. С. Будимирова

**Дизайн обложки:**  
А. В. Вожаков  
А. П. Сергеев

**Автор фото:**  
А. Г. Бувевич

**Адрес редакции:**  
620990, г. Екатеринбург,  
ИПЭ УрО РАН,  
ул. С. Ковалевской, 20  
Тел.: (343) 374-37-71  
E-mail: iie@ecko.uran.ru  
www.restrajjectory.ru

Журнал зарегистрирован  
Федеральной службой  
по надзору в сфере связи,  
информационных  
технологий и массовых  
коммуникаций  
27 апреля 2022 г.,  
Эл№ ФС77-83207

**Учредитель**  
**ИПЭ УрО РАН**

Перепечатка материалов,  
опубликованных в журнале,  
возможна только  
с разрешения редакции.

## ОТ ВЫПУСКАЮЩИХ РЕДАКТОРОВ

Дорогой читатель!

Находясь в самом начале третьего тысячелетия, мы с вами стали свидетелями изменений, вызванных деятельностью человека, скорость которых уже сейчас не позволяет за разумное время вполне осознать их последствия. Современная наука и современные технологии увеличивают нашу мощь явно быстрее, чем наделяют нас пониманием происходящего. Все осложняется тем, что процессы, происходящие в глобальной экосистеме, подобно торнадо, вовлекают во взаимодействие факторы совершенно различной природы, делая проблему «человек – среда обитания» не только глобальной, но и междисциплинарной. И эти процессы имеют склонность ускоряться, драматически сокращая горизонт предсказаний. Блуждая в пространстве возможностей, мы не знаем, когда пройдем точку невозврата, это может произойти за мгновение до катастрофы. Сказанное приводит нас к неизбежному выводу: исследования, направленные на моделирование и прогноз поведения сложных систем, невозможно переоценить.

Перед вами третий выпуск журнала «Траектория исследований – человек, природа, технологии», который посвящен вопросам математического моделирования в системе «человек – среда обитания».

Открывают выпуск четыре работы, связанные с медициной. Первая статья из этого списка посвящена методам визуализации данных в медико-биологических исследованиях – важному инструменту концентрированного представления результатов. Следующая статья знакомит читателя с подходом, основанным на скользящем среднем в линейной и нелинейной регрессии в биологии и медицине. Далее идут две работы, связанные с задачами реабилитации: одна посвящена методу управления экзоскелетом верхних конечностей, другая – алгоритму определения момента времени начала акта воображения движения для выявления паттернов нейронной активности.

Также в журнале можно познакомиться с двумя статьями о моделировании распределения примесей в компонентах среды обитания: в атмосферном воздухе и в верхнем слое почвы. Еще две работы носят обзорно-методический характер: одна из них – о широко применяемой в медико-биологических науках модели совместного действия бинарных факторов и равенстве рисков, другая – о методах анализа дискретных данных.

Завершает номер рубрика «Из научных архивов ИПЭ УрО РАН». В ней представлена статья основателя Института промышленной экологии УрО РАН, члена-корреспондента РАН Виктора Николаевича Чуканова, из сборника «Урал: наука, экология», опубликованного в 1999 году. Вступительное слово к этой статье написал давний коллега Виктора Николаевича, доктор физико-математических наук, профессор А. Н. Вараксин.

Е. Д. Константинова, А. П. Сергеев



# МЕТОДЫ ВИЗУАЛИЗАЦИИ ДАННЫХ В МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

С. Ю. Огородникова <sup>1</sup>, Е. Д. Константинова <sup>2</sup>

<sup>1</sup> Уральский Федеральный университет им. первого Президента России  
Б. Н. Ельцина, г. Екатеринбург, Россия

<sup>2</sup> Институт промышленной экологии УрО РАН, г. Екатеринбург, Россия

*В статье представлены результаты применения методов визуализации многомерных данных к решению прикладной задачи из области медицины. Используются ретроспективные данные 586 пациентов с раком мочевого пузыря, перенесших радикальную цистэктомию с декабря 2001 г. по май 2021 г. в онкоурологическом отделении Свердловского областного онкологического диспансера (СООД). В качестве методов визуализации многомерных данных реализованы такие подходы, как кривые выживаемости (по методу Каплана – Мейера), тепловые карты и деревья классификации. Для анализа данных использовались пакеты Statistica for Windows v.10. и MATLAB R2021b. На базе трех подходов показана возможность выявления отношений между численными данными на основе визуального анализа полученной совокупности графических образов. Продемонстрированы примеры построения наглядных, интуитивно понятных решающих правил для каждого из методов при решении задачи оценки 5- и 10-летней выживаемости пациентов Свердловского областного онкологического диспансера, перенесших радикальную цистэктомию тремя различными способами уродеривации. Рассмотрены особенности, преимущества и недостатки каждого из подходов применительно к работе с данными в медико-биологических исследованиях. Показано, что методы визуализации многомерных данных и подходы на их основе перспективны для анализа и наглядного представления экспериментальных численных результатов из области медицины.*

**Ключевые слова:** медико-биологические исследования; визуализация данных; тепловые карты; кривые выживаемости, деревья классификации.

## 1. Введение

Примерно с конца XVII столетия, с того момента, когда Эдмунд Галлей вывел формулы для расчета страховых платежей, статистика имеет дело с алгоритмами, помогающими в принятии решений [1]. За такой длительный срок масштабы собираемых данных претерпели существенные изменения – появились так называемые большие данные. Причем «большими» данные могут быть как по числу случаев (кейсов, строк), так и по количеству переменных (измеряемых параметров, характеристик, столбцов) в базе данных. Появление «больших данных» в медицине послужило стимулом для разработки алгоритмов и систем поддержки принятия ре-

шений в различных областях врачебной практики, включая такие ответственные, как оценка исходов оперативного вмешательства в онкологии.

Анализ медико-биологических данных (МБД) имеет ряд особенностей, которые переводят его из разряда рутинных задач в творческие. Выделим основные:

- разнообразие типов данных: количественные (непрерывные и дискретные), категориальные (номинальные, могут быть в бинарной форме), ранговые;
- загрязненность (или зашумленность) реальных МБД, часто требующих предварительной обработки;
- достаточно высокий уровень коррелированности реальных МБД;
- невозможность получения 100 %-й точности результатов моделирования, поскольку мы имеем дело с «паутиной причинности» возникновения патологии. Ни один фактор риска возникновения болезни сам по себе не является непосредственной причиной заболевания, он может лишь увеличить или изменить вероятность его появления;
- нежелательность применения моделей типа «черного ящика» с непрозрачной работой алгоритма, поскольку необходима контролируемость алгоритмов, потенциально влияющих на жизнь людей;
- желательность получения интуитивно понятных решающих правил модели для использования специалистами – медиками и биологами.

Две последние (по порядку, но не по значимости) особенности диктуют необходимость широкого внедрения в статистический анализ МБД методов *визуализации*.

Визуализацию данных можно определить как системное, основанное на правилах графическое представление информации, помогающее разобраться в сложных понятиях, нацеленное на обобщение, синтез теории и опыта [2].

Одна из целей визуализации – выделение закономерностей или аномалий в численных данных, а также первичная оценка набора данных для возможности применения в дальнейшем более сложных инструментов анализа. Применяя методы визуализации данных, исследователь нацелен на поиск наиболее выразительных изображений изучаемых объектов и связей между ними.

Хотелось бы особо подчеркнуть, что для решения одной прикладной задачи зачастую не существует единого подходящего метода визуализации. Это можно считать частным случаем известного принципа множественности моделей В. В. Налимова [3], который постулирует следующее: для объяснения и предсказания структуры и (или) поведения сложной системы возможно построение нескольких моделей, имеющих одинаковое право на существование. Применение комбинации нескольких методов для обеспечения высокого качества результатов является в данном случае необходимостью.

Цель работы – визуализировать численные данные из области медицины и показать возможность выявления отношений между ними на основе зрительного анализа полученной совокупности графических изображений (образов).

## 2. Материалы и методы

Клиническая часть работы по сбору данных выполнена в онкоурологическом отделении СООД. Были ретроспективно проанализированы данные 586 пациентов с раком мочевого пузыря, перенесших радикальную цистэктомию (РЦ) с декабря 2001 г. по май 2021 г. База данных включала в себя такие характеристики пациен-

тов, как пол, возраст, наличие хронических заболеваний, сведения о проведенной операции, наличие и тяжесть послеоперационных осложнений, а также данные о нескольких видах выживаемости.

В исследование были включены 523 мужчины (89,3 %) и 63 женщины (10,7 %) возрастом от 28 до 81 г. Средний возраст пациентов составил  $59,8 \pm 8,6$  лет.

Пациенты были разделены на 3 группы в зависимости от способа уродеривации, осуществленного в ходе операции:

- пациенты с наружным отведением мочи (группа 1,  $n = 82$  человека);
- пациенты с отведением мочи в изолированный сегмент подвздошной кишки (группа 2,  $n = 373$  человека);
- пациенты с ортоптическими кишечными резервуарами (группа 3,  $n = 131$  человек).

В статье рассмотрены и реализованы такие подходы визуализации многомерных данных, как кривые выживаемости, тепловые карты и деревья классификации.

Представленные в работе методы были выполнены в пакетах Statistica for Windows 10 version и MatLAB2021b.

На протокол исследования было получено одобрение Локального этического комитета ИПЭ УрО РАН.

### **3. Результаты**

Рассмотрим результаты решения прикладной задачи по оценке выживаемости пациентов СООД, перенесших РЦ тремя различными способами уродеривации.

#### **3.1. Кривые выживаемости**

Когда говорят об исследованиях в области визуализации информации, первым обычно вспоминают исследование «элементарных перцептивных задач» Уильяма Кливленда и Роберта МакГилла, которые еще в 1984 г. доказали, что сравнение объектов в одной шкале, например по оси, является самым простым визуальным действием [4].

Рассмотрим реализацию подхода, предложенного Капланом и Мейером в 1958 г. [5–7], позволяющего отображать многомерные объекты в виде кривых линий, достаточно простых и понятных для восприятия (рис. 1).



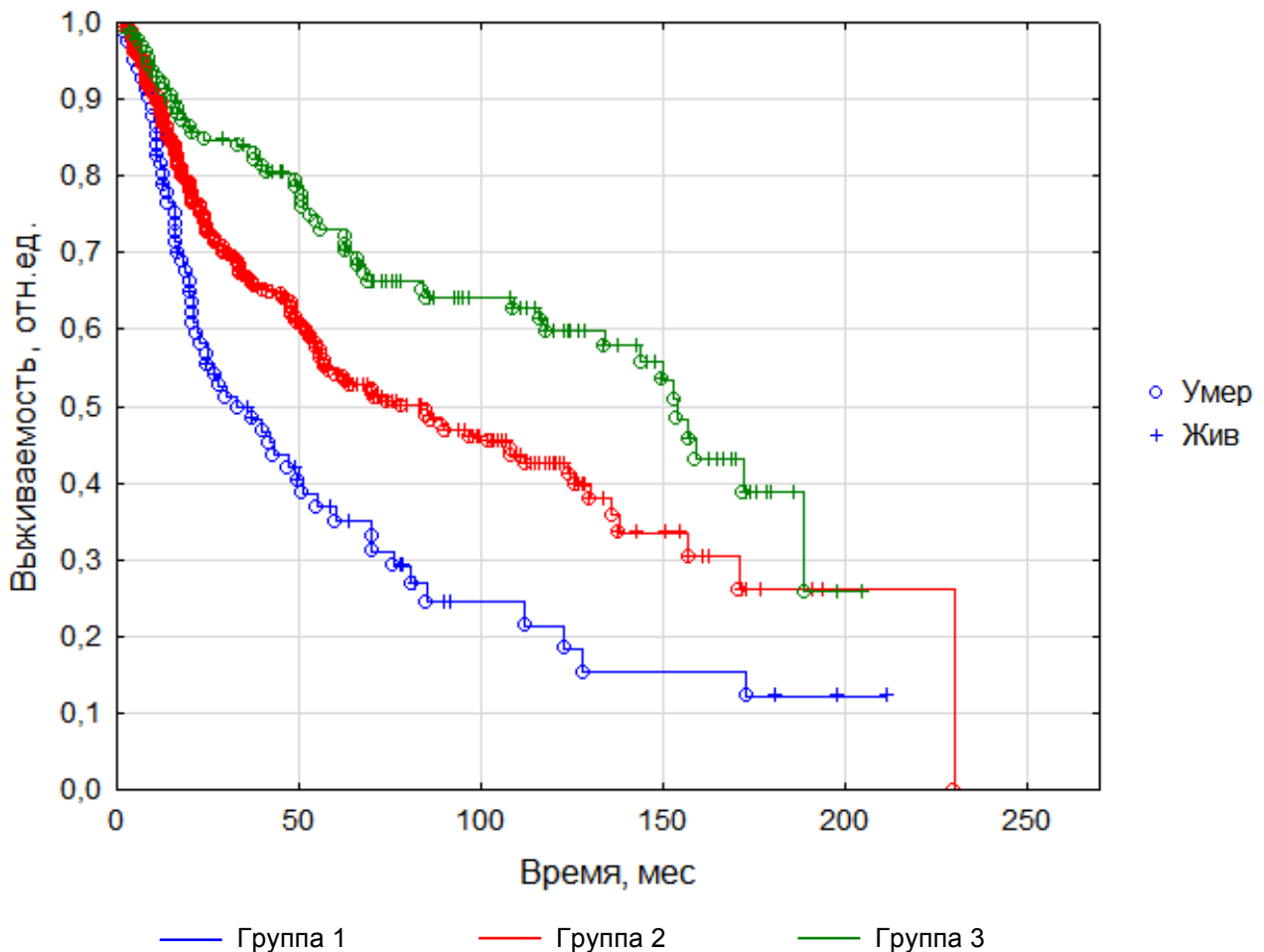


Рис. 1. Кривые общей выживаемости пациентов СООД, перенесших РЦ тремя способами уродеривации

Как следует читать график кривых выживаемости? По оси абсцисс отчается время в месяцах. Время наблюдения для живых пациентов определяют как время от момента постановки диагноза до даты последнего наблюдения за пациентом, а для умерших пациентов – как время от момента постановки диагноза до даты смерти пациента. По оси ординат – выживаемость в относительных единицах, где «1,0 = 100 % наблюдаемых пациентов живы», а «0,0 = 0 % все наблюдаемые пациенты умерли». Точка на кривой, отмеченная знаком «+» и «o», соответствуют пациенту, который на момент анализа данных был жив или мертв.

Визуальный анализ графика (рис. 1) позволяет сделать следующий вывод: выживаемость у пациентов трех групп различается (кривые достаточно удалены друг от друга). Лучшая выживаемость прослеживается у третьей группы больных, а худшая – у пациентов первой группы, по методу уродеривации.

Существует такой расчетный показатель, как медиана выживаемости ( $Me$ ) – значение времени, до которого выживаемость для выборки не опускалась ниже 0,5 (50 %). Стоит отдельно подчеркнуть, что речь идет именно о выживаемости, а не о доле выживших пациентов. Основное различие данных терминов заключается в том, что выживаемость представляет собой вероятность пережить определенный промежуток времени и учитывает выбывших пациентов. Доля выживших пациентов на момент времени  $t$  в свою очередь является отношением переживших момент  $t$  к объему выборки. В условиях отсутствия выбывших пациентов данные значения равны, однако чем больше будет выбывших, тем больше будет расхождение между этими значениями. Мы бы хотели сделать акцент на том, что в условиях реальных

медицинских исследований сложно представить ситуацию отсутствия выбывших из исследования (по разным причинам) пациентов.

Медианы общей выживаемости были рассчитаны для каждой группы: Ме в первой группе составила 37 месяцев, во второй – 78 месяцев, в третьей группе данный показатель оказался максимальным – 154 месяца. Рассчитанные показатели медиан выживаемости подтвердили результаты визуального анализа кривых – лучшую выживаемость демонстрируют пациенты третьей группы; худшую – пациенты первой группы, по методу уродеривации. При этом визуальный анализ интуитивно понятен, не требует специальной математической подготовки и проведения дополнительных расчетов.

### 3.2. Тепловые карты

Карты – один из древнейших способов визуализации, отображающий окружающую реальность. В различных областях научных исследований достаточно популярны тепловые карты – метод визуализации многомерных данных, разработанный и впервые примененный более 100 лет назад [8, 9].

Представленный подход реализован в пакете MatLAB2021b. На рис. 2 представлены тепловые карты различных видов выживаемости для трех групп пациентов СООД, перенесших РЦ.

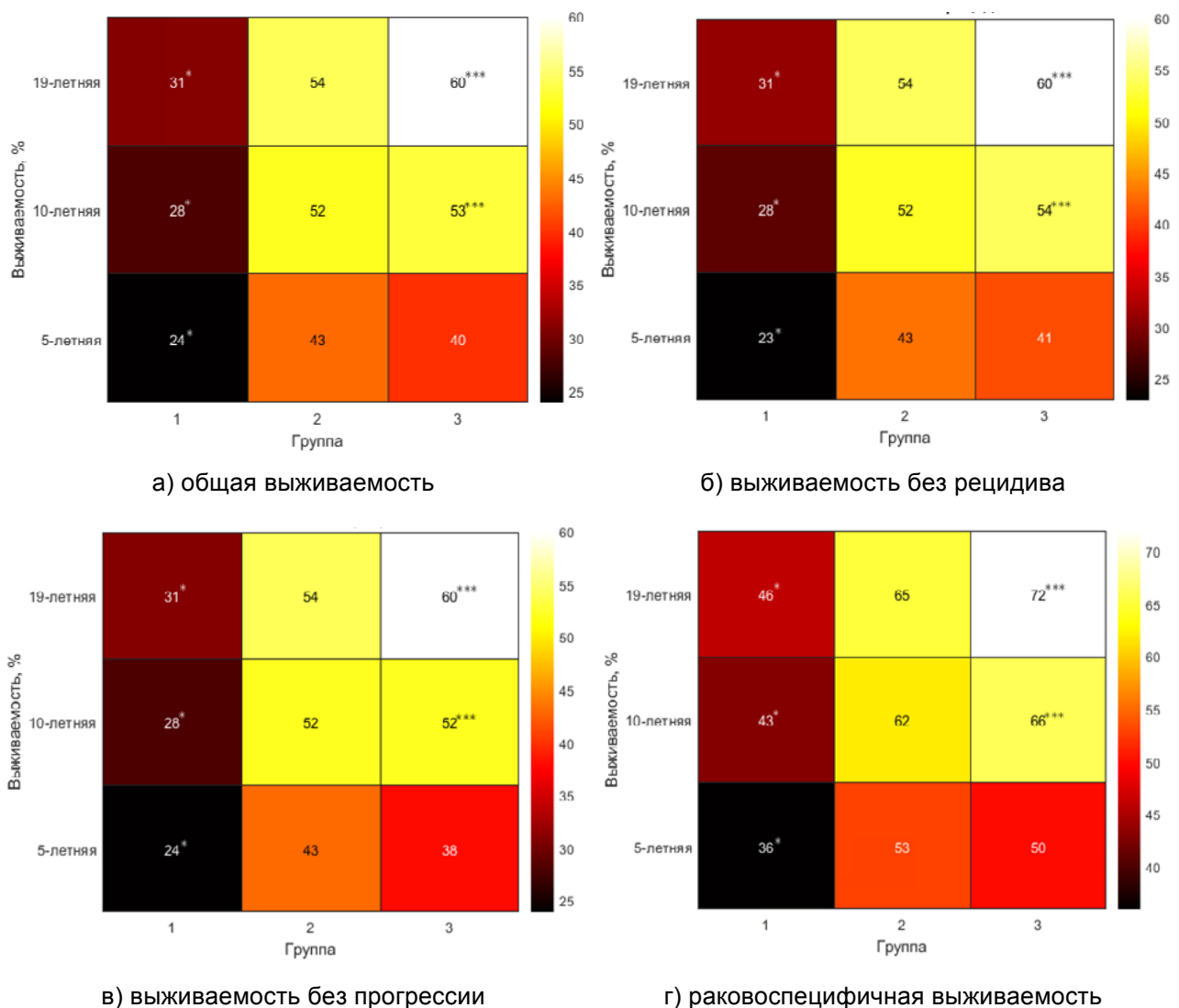


Рис. 2. Тепловые карты 5 и 10-летней выживаемости пациентов СООД

Интерпретацию тепловых карт проводят, опираясь на цветовую шкалу, расположенную справа от основного рисунка. В нашем исследовании использовалась наиболее удобная для восприятия последовательная цветовая шкала оттенков: от черного цвета (наименьшее значение выживаемости) к белому (наибольший процент выживших пациентов).

Рассмотрим в качестве примера общую 10-летнюю выживаемость пациентов 2 группы: нужная нам ячейка расположена на пересечении строки «10-летняя выживаемость» и столбца «2 группа». Значение выживаемости составляет 52 %; желтый цвет ячейки по последовательной цветовой шкале оттенков близок к белому, что соответствует лучшим показателям выживаемости. Соседняя ячейка (10-летняя выживаемость пациентов 1 группы) окрашена в более темный оттенок, что говорит о худших (относительно 2 группы) показателях выживаемости (28 %).

Анализируя полученные карты (рис. 2), можем говорить об одинаковом характере поведения всех рассмотренных видов выживаемости для трех групп пациентов: наименьшая выживаемость у пациентов первой группы на всех временных промежутках (левый столбец), для пациентов второй и третьей групп показатели выживаемости практически не различаются, наибольшее значение (60 и 72 месяца) соответствует выживаемости пациентов третьей группы за все время исследования (19 лет).

### 3.3. Деревья классификации

Одна из ответственных областей применения статистической обработки медицинских данных – это оценка исходов оперативного вмешательства. Большое количество данных, влияющих на выбор лечения пациента, вызывает необходимость разработки алгоритмов и систем поддержки принятия врачебных решений. При этом главными требованиями к подобным алгоритмам остаются прозрачность, наглядность и простота интерпретации решающих правил. Наиболее ярко отражает черты многомерного анализа данных, отвечая при этом всем вышеупомянутым требованиям, алгоритм построения деревьев классификации (деревьев решений).

Были построены деревья решений для оценки 5- и 10-летней выживаемости пациентов СООД, перенесших РЦ (рис. 3, 4).

Представленный подход реализован в пакете Statistica for Windows v.10., по описанной в [10–12] методике.



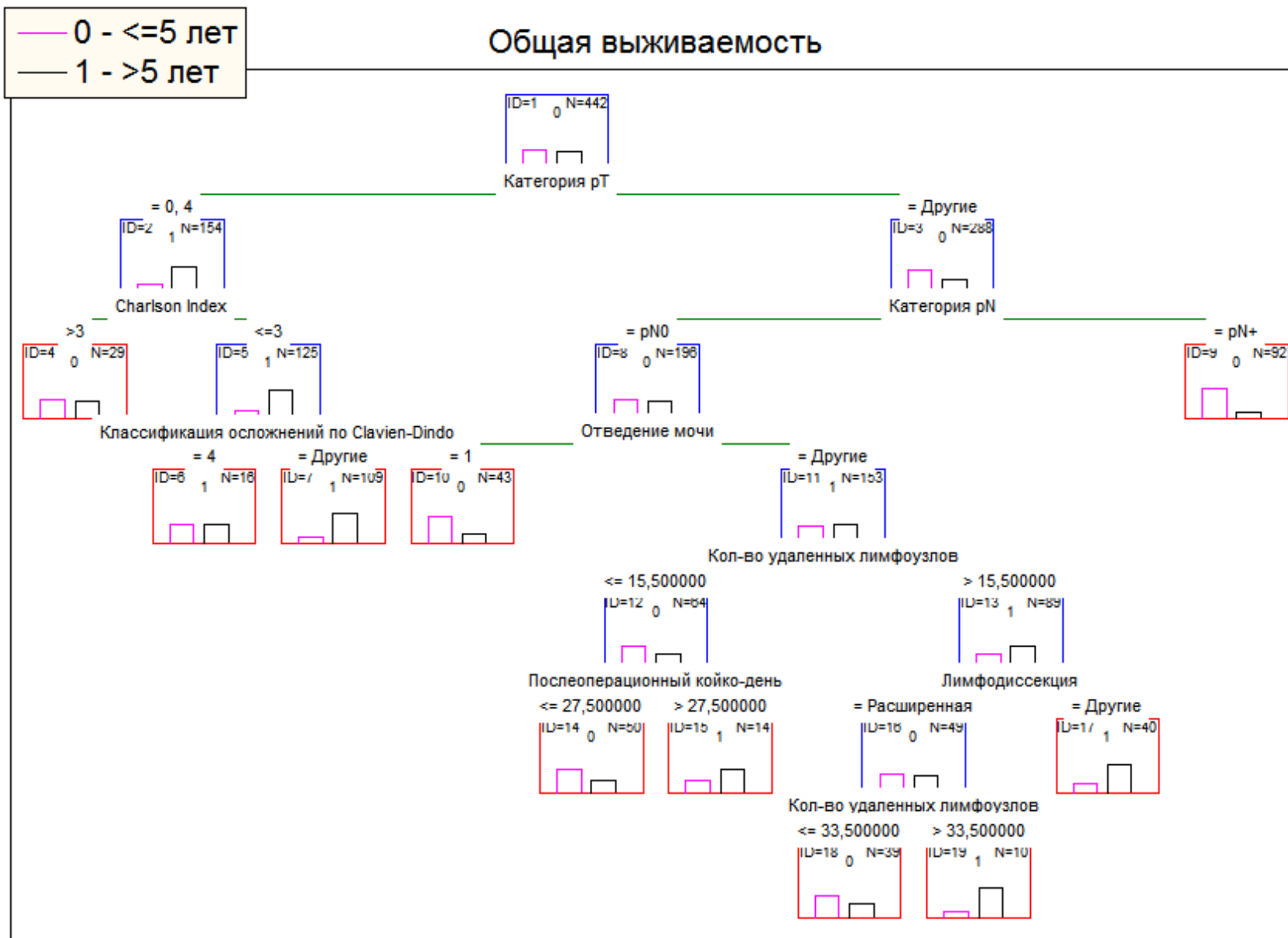


Рис. 3. Дерево решений для оценки 5-летней выживаемости



На рис. 3 представлено дерево решений для оценки 5-летней выживаемости пациентов СООД, перенесших РЦ, построенное методом ветвления CART с оцениваемыми априорными вероятностями, равными ценами ошибок, и фактической остановкой алгоритма с долей неклассифицированных объектов 0,1. Представленная модель содержит 9 терминальных вершин. Чувствительность такой классификации составляет 80,3 %, специфичность – 71,9 %, а точность – 76,0 %.

Дерево решений для оценки 10-летней выживаемости пациентов СООД, перенесших РЦ (рис. 4), построено методом ветвления CART с оцениваемыми априорными вероятностями, равными ценами ошибок, и остановкой алгоритма после первой ошибки классификации. Представленная модель содержит 8 терминальных вершин, чувствительность такой классификации составляет 37,5 %, специфичность – 93,5 %, а точность – 80,3 %.

На примере рис. 4 опишем, как трактовать представленное на нем дерево классификации.

Самая верхняя вершина носит название корневой и содержит всю исследуемую выборку. Алгоритм делит корневую вершину (исходную выборку) на две промежуточные. Название переменной, по которой происходит деление (ветвление), указано под разделяемой вершиной. Последовательно задаются вопросы, ответы на которые могут принимать только два значения. Окончательное решение зависит от ответов на все предыдущие вопросы. Процесс ветвления продолжается до остановки по выбранному критерию. Вершины, после которых не производится дальнейшее ветвление, называются терминальными.

Проследим, как формируется класс пациентов, которые смогут прожить 10 и более лет после проведения оперативного вмешательства (класс «1»). Для этого пройдем путь от корневой вершины до терминальных вершин, помеченных «1». Решающее правило попадания в класс «1» формируется, по сути, описанием этого маршрута. Пациент сможет прожить 10 и более лет после проведения оперативного вмешательства, если он из 1-й или 2-й группы по способу уродеривации; с кровопотерей в ходе операции 425 мл или меньше; при любой категории рТ, кроме 4-й; и с низкой степенью злокачественности опухоли (G1). Описание решающего правила метода ДК похоже на то, как положение листа на дереве можно описать, указав ведущую к нему последовательность ветвей (начиная со ствола и заканчивая самой последней веткой, на которой располагается лист).

## 4. Обсуждение

Основная цель любой информационной технологии – это получение исследователем адекватной информации для ее дальнейшего анализа и принятия на его основе какого-либо решения. Мы показали, как методы визуализации позволяют буквально «одним взглядом» обнаружить особенности и выявить закономерности в больших объемах экспериментальных численных данных.

Перейдем к подробному обсуждению каждого из представленных подходов.

Такой подход визуализации многомерных данных, как *кривые выживаемости по методу Каплана – Мейера*, широко применяется в медицинских исследованиях и является золотым стандартом при оценке эффективности лечения в онкологии. Метод реализуем во многих современных пакетах статистической обработки данных (Statistica, SPSS, SAS и др.).

Интерпретируя *кривые выживаемости*, стоит помнить, что для оценки статистической значимости различий одного визуального анализа недостаточно. Возни-



кает необходимость в применении математического алгоритма сравнения кривых выживаемости, результатом которого является вычисленное  $p$ -значение, по близости к уровню значимости  $\alpha$  которого возможно судить о статистически значимых различиях между полученными кривыми. Существует несколько математических алгоритмов, успешно зарекомендовавших себя при определенных условиях поставленной задачи. В медицинских исследованиях сравнение кривых выживаемости проводят с помощью Log-Rank Test или логрангового критерия [13–16].

Как было показано в разделе «Результаты», для того чтобы сделать вывод о статистической значимости различий, можно сравнивать и медианы выживаемости. Сравнение происходит по аналогии с логранговым критерием, с использованием так называемого медианного теста, являющегося частным случаем критерия Краскела – Уоллиса и непараметрической альтернативой одномерному дисперсионному анализу.

Широкое распространение и признание метод Каплана – Мейера получил благодаря прозрачному и простому для понимания алгоритму построения кривых выживаемости.

Относительным ограничением распространенности метода, особенностью применения подхода во врачебной практике, можно считать необходимость в дополнительной процедуре сравнения кривых выживаемости с помощью статистических критериев, что требует определенной математической подготовки.

Следующим рассмотренным подходом были *тепловые карты*, которые относятся к методам визуализации многомерных данных с помощью цветных индикаторов. При этом табличное представление информации все же сохраняется, но числа в ячейках заменяются на заливку цветом согласно выбранной шкале.

Важным фактором для правильной и интуитивно понятной интерпретации тепловой карты является хорошо подобранная цветовая схема. Существуют расходящиеся и последовательные цветовые схемы [17]. Расходящиеся палитры фиксируют цвета в нижнем и верхнем конце, а также в середине диапазона данных. Такое решение лучше подходит для данных, которые варьируются как в отрицательном, так и в положительном направлениях. Последовательные же палитры фиксируют наименьшее и наибольшее значения всего диапазона данных, оптимальны для неотрицательных данных (например, процент от 0 до 100). В нашем исследовании использована последовательная палитра, удобная для восприятия при отсутствии в анализе отрицательных численных данных. Этот метод представления данных активно применяется в исследованиях в области медицины, экономики и социологии [18–22].

В контексте применения во врачебной практике к преимуществам *тепловых карт* как метода визуализации многомерных данных можно отнести прозрачность алгоритма, наглядность результатов анализа, а также простоту интерпретации полученных графических образов.

Из недостатков тепловых карт можно выделить сложность обработки слишком большого объема данных, так как при этом метод теряет свое основное преимущество – простоту интерпретации. Общая схема такой карты слишком массивна, а отдельные ячейки в ней плохо различимы, что приводит к возможности оценить лишь «общую картину» происходящего.

Последним по порядку изложения, но отнюдь не по значению, является метод *деревьев классификации (ДК)*. Применение метода ДК, на наш взгляд, способствует быстрому выявлению скрытых связей между данными в самой очевидной форме.

К преимуществам метода можно отнести высокую интерпретируемость модели, поддержку и числовых, и категориальных признаков, малое число входящих параметров. Легко визуализируется как само дерево, так и конкретное решающее правило (путь в дереве от корневой вершины к терминальной). К тому же алгоритм ДК способен учитывать опыт группы экспертов, рассматривающих конкретную проблему. Последнее обстоятельство повышает актуальность предложенного подхода.

К недостаткам метода ДК, построенных способом CART, можно отнести нестабильность. Небольшие изменения в данных могут существенно изменять построенное дерево решений. Эту проблему решают с помощью построения ансамблей деревьев классификации [3, 10–12].

Анализируя показатели качества построенных нами деревьев решений (см. раздел «Результаты»), нетрудно заметить, что модели качественно классифицируют лишь отрицательные случаи (пациенты СООД, не доживающие до 5- и 10-летней временной отметки). В ряде случаев такие модели допускаются к практическому применению, так как задача подразумевает точное определение только одной группы классифицированных объектов. Например, при построении прогностической модели оценки платежеспособности потенциальных заемщиков более важной задачей является выявление и классификация кредиторов с потенциальными задолженностями, нежели порядочных заемщиков, исправно вносящих платежи.

В случае же медицинских исследований такую модель не всегда можно назвать вполне удовлетворительной, так как ошибочная классификация «здоровых» пациентов в группу риска может спровоцировать ухудшение качества их жизни. Однако это не повод усложнять дерево, добавляя новые вершины в погоне за высоким процентом правильной классификации. Стоит признать, что при работе с МБД количественная эффективность не может быть принята как единственный критерий качества алгоритма, и ради сохранения простоты алгоритма бывает разумно отказаться от попыток достичь более высоких показателей его эффективности. Это обстоятельство еще раз подчеркивает необходимость совместной работы с экспертами в данной области.

## 5. Заключение

На базе трех подходов показана возможность выявления отношений между численными данными на основе визуального анализа полученной совокупности графических образов. Тот факт, что результаты анализа доступны для понимания медикам, а не только специалистам в математике и статистике, является большим преимуществом использования методов визуализации.

Определенно, за применением графики в исследованиях большое будущее, и не только потому, что увеличивается скорость передачи информации и повышается уровень ее понимания. Возможно, гораздо важнее тот факт, что применение методов визуализации способствует развитию таких важных для любого исследователя качеств, как интуиция и образное мышление.

## 6. Выводы

1. Показано, что методы визуализации многомерных данных и подходы на их основе перспективны для анализа и наглядного представления экспериментальных численных результатов из области медицины.
2. Для трех групп пациентов СООД, перенесших РЦ, построены кривые выживаемости методом Каплана – Мейера. Визуальный анализ кривых подтвердил

результат, полученный путем расчета медиан выживаемости. При этом визуальный анализ прост для понимания, не требует специальной математической подготовки и проведения дополнительных расчетов.

3. Построены тепловые карты для оценки 5- и 10-летней выживаемости пациентов СООД, перенесших РЦ тремя различными способами уродеривации. Продемонстрированы очевидные преимущества метода: максимально простой и понятный для восприятия результат в виде таблицы, ячейки которой окрашены согласно подобранной цветовой шкале. Данный подход оптимален в решении задач, которые требуют общего заключения, а также как первичная оценка набора данных для возможности применения в дальнейшем более сложных инструментов анализа.
4. Построены деревья решений для оценки 5- и 10-летней выживаемости пациентов, перенесших РЦ тремя различными способами уродеривации. Продемонстрировано, что алгоритмы на основе ДК способны не только наглядно визуализировать сложные связи между набором переменных, но и послужить основой для построения прогностических моделей оценки эффективности лечения.
5. Показано, что для объяснения поведения сложной системы возможно построение нескольких моделей, имеющих одинаковое право на существование.

## 7. Список литературы

1. *Шпигельхалтер, Д.* Искусство статистики. Как находить ответы в данных / Д. Шпигельхалтер. – М. : Манн, 2021. – 449 с. – ISBN 978-5-00-169250-8.
2. Применение методов визуализации при исследовании структуры экспериментальных многомерных данных / В. А. Воловоденко, О. Г. Берестнева, Е. В. Немеров, И. А. Осадчая // Известия Томского политехнического университета, 2012. – Т. 320, № 5. – С. 125–130. – EDN OZQTGT.
3. *Налимов, В. В.* Теория эксперимента / В. В. Налимов. – М. : Наука, 1971. – 206 с.
4. *Cleveland, W. S.* Graphical perception: Theory, experimentation, and application to the development of graphical methods / W. S. Cleveland, R. McGill // J. of the American statistical association. – 1984. – Т. 79, No. 387. – P. 531–554. – DOI 10.1080/01621459.1984.10478080.
5. *Гланц, С.* Медико-биологическая статистика / С. Гланц. – М. : Практика, 1999. – 459 с.
6. Estimating hazard ratios from published Kaplan – Meier survival curves: A methods validation study / R. Saluja, S. Cheng, K. A. delos Santos, K. K. Chan // Research Synthesis Methods. – 2019. – Т. 10, No. 3. – P. 465–475. – DOI: 10.1002/jrsm.1362.
7. *Messori, A.* Synthetizing published evidence on survival by reconstruction of patient-level data and generation of a multi-trial Kaplan – Meier curve / A. Messori // Cureus. – 2021. – Т. 13, No. 11. – DOI: 10.7759/cureus.19422.
8. *Wilkinson, L.* The history of the cluster heat map / L. Wilkinson, M. Friendly // The American Statistician. – 2009. – Т. 63, No. 2. – P. 179–184. – DOI: 10.1198/tas.2009.0033.
9. *Романова, И. К.* Современные методы визуализации многомерных данных: анализ, классификация, реализация, приложения в технических системах / И. К. Романова // Наука и образование: научное издание МГТУ им. Н. Э. Баумана. – 2016. – № 3. – С. 133–167. – DOI 10.7463/0316.0834876. – EDN VTKQIZ.
10. *Buntine, W.* Learning classification trees / W. Buntine // Artificial Intelligence frontiers in statistics. – 2020. – P. 182–201.

11. Benchmarking cesarean delivery rates using machine learning-derived optimal classification trees / A. C. Gimovsky, D. Zhuo, J. T. Levine [et al.] // Health Services Research. – 2022. – Т. 57, No. 4. – P. 796–805. – DOI: 10.1111/1475-6773.13921.
12. *Нессонова, М. Н.* Математические модели и методы построения классификаторов в медицине / М. Н. Нессонова. – LAP LAMBERT Academic Publishing, 2018. – 213 с. – ISBN 987-613-9-58671-4.
13. Does radical cystectomy improve overall survival in octogenarians with muscle-invasive bladder cancer? / S. Yoo, D. You, I. G. Jeong [et al.] // Korean J. of Urology. – 2011. – Т. 52, No. 7. – P. 446–451. – DOI: 10.4111/kju.2011.52.7.446.
14. Long-term oncologic outcomes after radical cystectomy for bladder cancer at a single institution / T. Kwon, I. G. Jeong, D. You [et al.] // J. of Korean medical science. – 2014. – Т. 29, No. 5. – P. 669–675. – DOI: 10.3346/jkms.2014.29.5.669.
15. *Schober, P.* Survival analysis and interpretation of time-to-event data: the tortoise and the hare / P. Schober, T. R. Vetter // Anesthesia and analgesia. – 2018. – Т. 127, No. 3. – P. 792. – DOI: 10.1213/ANE.0000000000003653.
16. Opencrimemapping. org: An Online Tool for Visualizing Crime / M. Crowder, L. Darr, G. Garza, B. Allen // SMU Data Science Review. – 2018. – Т. 1, No. 3. – P. 11.
17. *Harrower, M.* Colorbrewer.org: an online tool for selecting colour schemes for maps / M. Harrower, C. A. Brewer // The Cartographic J. – 2003. – Т. 40, No. 1. – P. 27–37.
18. *Жгун, Т. В.* Визуализация данных при оценке изменения качества социально-экономических систем / Т. В. Жгун, П. Д. Кристофер // Системный анализ в проектировании и управлении. – 2020. – Т. 24, № 1. – С. 257–267. – DOI: 10.18720/SPBPU/2/id20-130.
19. *DeBold, T.* Battling infectious diseases in the 20<sup>th</sup> century: the impact of vaccines / T. DeBold, D. Friedman // Wall Street J. – 2015. – URL: <http://graphics.wsj.com/infectious-diseases-and-vaccines/>.
20. Heat map visualization for electrocardiogram data analysis / H. Guo, W. Zhang, C. Ni [et al.] // BMC cardiovascular disorders. – 2020. – Т. 20, No. 1. – P. 1–8. – DOI: 10.1186/s12872-020-01560-8.
21. *Metsalu, T.* ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap / T. Metsalu, J. Vilo // Nucleic acids research. – 2015. – Т. 43, No. W1. – P. 566–570. – DOI: 10.1093/nar/gkv468.
22. *Lee, S. Y.* High-dose drug heat map analysis for drug safety and efficacy in multi-spheroid brain normal cells and GBM patient-derived cells / S. Y. Lee, Y. Teng, M. Son [et al.] // Plos one. – 2021. – Т. 16, No. 12. – e0251998. – DOI: 10.1371/journal.pone.0251998.

## Сведения об авторах:

**Огородникова Светлана Юрьевна**, студентка Уральского федерального университета им. первого Президента России Б. Н. Ельцина, г. Екатеринбург, Россия. Эл. почта: [ogorodnikova.sveta2507@gmail.com](mailto:ogorodnikova.sveta2507@gmail.com).

**Константинова Екатерина Даниловна**, канд. ф.-м. н., старший научный сотрудник, и. о. зав. лаборатории Математического моделирования в экологии и медицине Института промышленной экологии УрО РАН, г. Екатеринбург, Россия. Эл. почта: [K\\_Konst@ecko.uran.ru](mailto:K_Konst@ecko.uran.ru)

## DATA VISUALIZATION METHODS IN BIOMEDICAL RESEARCH

S. Y. Ogorodnikova <sup>1</sup>, E. D. Konstantinova <sup>2</sup>

<sup>1</sup> Ural Federal University named after the first President of Russia B. N. Yeltsin,  
Ekaterinburg, Russia

<sup>2</sup> Institute of Industrial Ecology, Ural Branch, Russian Academy of Sciences,  
Ekaterinburg, Russia

*The article presents the results of applying multidimensional data visualization methods to solving an applied problem from the field of medicine. We used retrospective data of 586 patients with bladder cancer who underwent radical cystectomy from December 2001 to May 2021 in the urological oncology department of the Sverdlovsk Regional Oncology Center. Such approaches as survival curves (according to the Kaplan – Meier method), heat maps and classification trees are implemented as methods for visualizing multidimensional data. The Statistica for Windows 10 version and MatLAB2021b packages were used for data analysis. On the basis of three approaches, the possibility of identifying relationships between numerical data based on a visual analysis of the resulting set of graphic images is shown. Examples of constructing visual, intuitive decision rules for each of the methods in solving the problem of assessing the 5- and 10-year survival of Sverdlovsk Regional Oncology Center patients who underwent radical cystectomy with three different methods of uroderivation are demonstrated. The features, advantages and disadvantages of each of the approaches in relation to working with data in biomedical research are considered. It is shown that the methods of visualization of multidimensional data and approaches based on them are promising for the analysis and visual presentation of experimental numerical results from the field of medicine.*

**Key words:** biomedical research; data visualization; heat maps; survival curves; classification trees.

### References

1. *Shpigel'halter, D.* The art of statistics. How to find answers in data / D. Shpigel'halter. – M. : Mann, 2021. – 449 p. – ISBN 978-5-00-169250-8. (In Russian).
2. Application of visualization methods in the study of the structure of experimental multi-dimensional data / V. A. Volovodenko, O. G. Berestneva, E. V. Nemerov, I. A. Osadchaya // Proceedings of the Tomsk Polytechnic University, 2012. – Vol. 320, No. 5. – P. 125–130. – EDN OZQTGT. (In Russian).
3. *Nalimov, V. V.* Theory of experiment / V. V. Nalimov. – M. : Nauka, 1971. – 206 p. (In Russian).
4. *Cleveland, W. S.* Graphical perception: Theory, experimentation, and application to the development of graphical methods / W. S. Cleveland, R. McGill // J. of the American statistical association. – 1984. – T. 79, No. 387. – P. 531–554. – DOI 10.1080/01621459.1984.10478080.
5. *Glants, S.* Medico-biological statistics / S. Glants. – M. : Praktika, 1999. – 459 p. (In Russian).

6. Estimating hazard ratios from published Kaplan – Meier survival curves: A methods validation study / R. Saluja, S. Cheng, K. A. delos Santos, K. K. Chan // *Research Synthesis Methods*. – 2019. – Т. 10, No. 3. – P. 465–475. – DOI: 10.1002/jrsm.1362.
7. *Messori, A.* Synthetizing published evidence on survival by reconstruction of patient-level data and generation of a multi-trial Kaplan – Meier curve / A. Messori // *Cureus*. – 2021. – Т. 13, No. 11. – DOI: 10.7759/cureus.19422.
8. *Wilkinson, L.* The history of the cluster heat map / L. Wilkinson, M. Friendly // *The American Statistician*. – 2009. – Т. 63, No. 2. – P. 179–184. – DOI: 10.1198/tas.2009.0033.
9. *Romanova, I. K.* Modern methods of visualization of multidimensional data: analysis, classification, implementation, applications in technical systems / I. K. Romanova // *Science and Education: scientific edition of Bauman Moscow State Technical University*. – 2016, No. 3. – P. 133–167. – DOI 10.7463/0316.0834876. – EDN VTKQIZ. (In Russian).
10. *Buntine, W.* Learning classification trees / W. Buntine // *Artificial Intelligence frontiers in statistics*. – 2020. – P. 182–201.
11. Benchmarking cesarean delivery rates using machine learning-derived optimal classification trees / A. C. Gimovsky, D. Zhuo, J. T. Levine [et al.] // *Health Services Research*. – 2022. – Т. 57, No. 4. – P. 796–805. – DOI: 10.1111/1475-6773.13921.
12. *Nessonova, M. N.* Mathematical models and methods of constructing classifiers in medicine / M. N. Nessonova. – LAP LAMBERT Academic Publishing, 2018. – 213 p. – ISBN 987-613-9-58671-4. (In Russian).
13. Does radical cystectomy improve overall survival in octogenarians with muscle-invasive bladder cancer? / S. Yoo, D. You, I. G. Jeong [et al.] // *Korean J. of Urology*. – 2011. – Т. 52, No. 7. – P. 446–451. – DOI: 10.4111/kju.2011.52.7.446.
14. Long-term oncologic outcomes after radical cystectomy for bladder cancer at a single institution / T. Kwon, I. G. Jeong, D. You [et al.] // *J. of Korean medical science*. – 2014. – Т. 29, No. 5. – P. 669–675. – DOI: 10.3346/jkms.2014.29.5.669.
15. *Schober, P.* Survival analysis and interpretation of time-to-event data: the tortoise and the hare / P. Schober, T. R. Vetter // *Anesthesia and analgesia*. – 2018. – Т. 127, No. 3. – P. 792. – DOI: 10.1213/ANE.0000000000003653.
16. *OpenCrimemapping.org: An Online Tool for Visualizing Crime* / M. Crowder, L. Darr, G. Garza, B. Allen // *SMU Data Science Review*. – 2018. – Т. 1, No. 3. – P. 11.
17. *Harrower, M.* Colorbrewer.org: an online tool for selecting colour schemes for maps / M. Harrower, C. A. Brewer // *The Cartographic J.* – 2003. – Т. 40, No. 1. – P. 27–37.
18. *Zhgun, T. V.* Data visualization in assessing changes in the quality of socio-economic systems / T. V. Zhgun, P. D. Christopher // *System analysis in design and management*. – 2020. – Vol. 24, No. 1. – P. 257–267. – DOI: 10.18720/SPBPU/2/id20-130.
19. *DeBold, T.* Battling infectious diseases in the 20<sup>th</sup> century: the impact of vaccines / T. DeBold, D. Friedman // *Wall Street J.* – 2015. – URL: <http://graphics.wsj.com/infectious-diseases-and-vaccines/>.
20. Heat map visualization for electrocardiogram data analysis / H. Guo, W. Zhang, C. Ni [et al.] // *BMC cardiovascular disorders*. – 2020. – Т. 20, No. 1. – P. 1–8. – DOI: 10.1186/s12872-020-01560-8.
21. *Metsalu, T.* ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap / T. Metsalu, J. Vilo // *Nucleic acids research*. – 2015. – Т. 43, No. W1. – P. 566–570. – DOI: 10.1093/nar/gkv468.
22. *Lee, S. Y.* High-dose drug heat map analysis for drug safety and efficacy in multi-spheroid brain normal cells and GBM patient-derived cells / S. Y. Lee, Y. Teng, M. Son [et al.] // *Plos one*. – 2021. – Т. 16, No. 12. – e0251998. – DOI: 10.1371/journal.pone.0251998.



# ЛИНЕЙНАЯ И НЕЛИНЕЙНАЯ РЕГРЕССИЯ В БИОЛОГИИ И МЕДИЦИНЕ: ПОДХОД, ОСНОВАННЫЙ НА СКОЛЬЗЯЩЕМ СРЕДНЕМ

А. Н. Варакин, Ю. В. Шалаумова, Т. А. Маслакова

*Институт промышленной экологии УрО РАН, г. Екатеринбург, Россия*

*В методической статье обсуждаются условия применимости линейной и нелинейной (логистической) регрессии. В качестве одного из методов проверки линейности связи между предиктором  $X$  и откликом  $Y$  в линейной регрессии и линейности связи предиктора  $X$  с логитом отклика  $Y$  в логистической регрессии предлагается подход, основанный на скользящем среднем. На конкретных примерах показаны преимущества скользящего среднего по сравнению с обычной стратификацией данных. Подчеркивается важность графического представления результатов скользящего среднего в линейной и (особенно) в логистической регрессии.*

**Ключевые слова:** скользящее среднее; проверка линейности; линейная регрессия; логистическая регрессия.

## 1. Введение

Одним из условий применимости регрессионных моделей является линейность связи между предиктором  $X$  и откликом  $Y$  в линейной регрессии и линейность связи предиктора  $X$  с логитом отклика  $Y$  в логистической регрессии. Существуют разнообразные методы для проверки линейности связи, основанные как на визуальной оценке графического представления данных, так и на вычислительных процедурах [1, 2]. Однако далеко не всегда в научных исследованиях действительно проводится процедура проверки линейности, что может оказать влияние на достоверность выводов. Если нарушения условий применимости не устраняются должным образом, то результаты могут быть недостоверными, в частности, может снижаться эффективность статистических тестов и увеличиваться частота ошибок I и II рода [3].

Если обратиться к истории вопроса, то предпосылкой к появлению регрессионного анализа стала разработка метода наименьших квадратов, опубликованного А. М. Лежандром в 1805 г. и К. Ф. Гауссом в 1809 г. для определения орбит тел в Солнечной системе. Пионерной публикацией, включавшей термин «регрессия», стали материалы доклада Ф. Гальтона в журнале *Nature* 1885 г., описывающие регрессионную связь между ростом отцов и сыновей [4]. Более формальная трактовка этого понятия впервые была сделана К. Пирсоном в работе 1896 г. [5]. Современный регрессионный анализ в значительной мере базируется на публикациях Р. А. Фишера 1920-х гг., объединившего работы Гаусса и Пирсона [6]. После Фишера было разработано множество важных расширений регрессии, включая ло-

гистическую, непараметрическую, байесовскую регрессию и др. [4]. Как для линейной, так и для логистической регрессии одним из ключевых условий применимости статистической модели является линейность [1].

Данная работа посвящена описанию метода проверки линейности связи в линейных и логистических регрессионных моделях посредством подхода, основанного на скользящем среднем.

## 2. Линейная регрессия

Линейная регрессия применяется для описания статистической связи между двумя количественными переменными  $Y$  и  $X$ . Переменная  $Y$  в линейной регрессии называется объясняемой переменной, значения которой объясняются предсказывающей переменной  $X$ . Переменную  $Y$  называют еще откликом (outcome) на воздействие предикторной (предсказывающей) переменной  $X$ . Предикторную переменную  $X$  для краткости называют «предиктор» [7, 8].

В первоначальной постановке задачи регрессионного анализа предполагалось, что предиктор  $X$  принимает ряд фиксированных значений, для каждого из которых имеется некоторое множество значений отклика  $Y$  (рис. 1а); это так называемый *fixed-X case* [9]. Именно для такого случая регрессионная модель формулировалась в виде:

$$E(Y|X) = b_0 + b_1X, \quad (1)$$

где символ  $E$  означает математическое ожидание значений  $Y$ , принадлежащих данному  $X$ . В практической работе с конкретными данными операция математического ожидания означает расчет среднего значения  $Y$  при заданном  $X$ .

Ситуация, изображенная на рис. 1а, чаще соответствует случаю планируемых экспериментов, когда исследователь имеет возможность задавать конкретные фиксированные значения предиктора  $X$ , исходя из поставленной задачи. Это область биологических экспериментов или клинических испытаний [11–13]. Совсем другая ситуация имеет место при проведении так называемых обсервационных исследований, когда исследователь использует существующие данные, не имея возможности их изменять или на них влиять [14]. К этой категории относятся натурные исследования в биологии или эпидемиологические исследования в медицине [15, 16]. При анализе результатов обсервационных исследований данные для линейного регрессионного анализа имеют вид, изображенный на рис. 1б: предиктор  $X$  принимает произвольный ряд значений, и для каждого значения  $X$  имеется (как правило) одно значение отклика  $Y$ , это так называемый *variable-X case* [9].

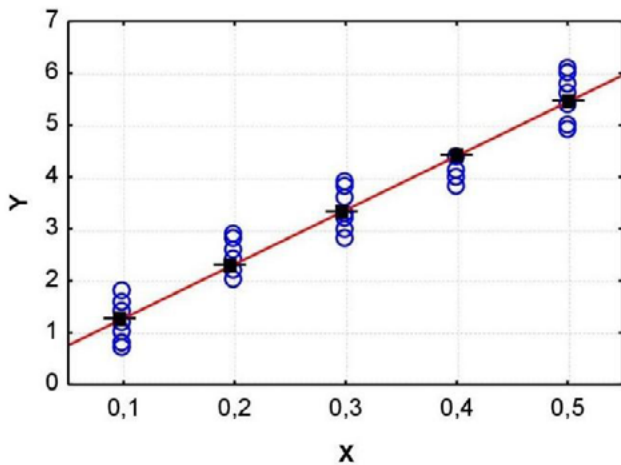


Рис. 1а. Пример зависимости  $Y(X)$  для планируемого эксперимента. Кружки – отдельные наблюдения, квадраты – средние значения  $Y$  для данного  $X$

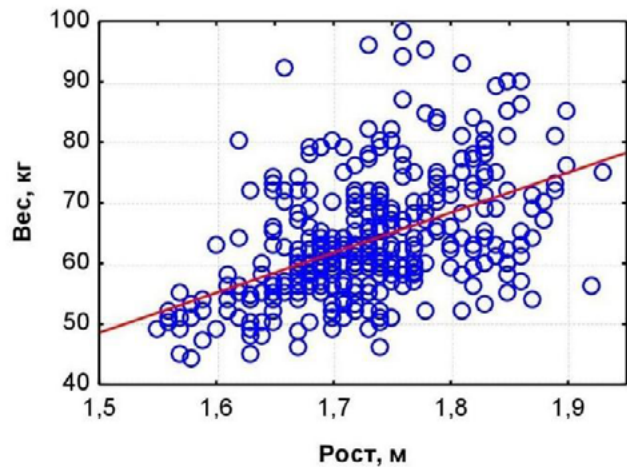


Рис. 1б. Соотношение веса и роста подростков [10]

Линейная регрессия (1) может быть использована только при выполнении ряда условий (условия применимости линейной регрессии):

А) средние значения отклика  $Y$  линейно связаны со значениями  $X$ ;

Б) распределение значений  $Y$  для каждого  $X$  подчиняется нормальному закону  $N(\mu_i, \sigma^2)$  со значениями математического ожидания  $\mu_i$  для значений предиктора  $X_i$  и с одинаковыми дисперсиями  $\sigma^2$  для каждого значения предиктора  $X_i$  [7, 8, 17].

Условие (А) относится к применимости модели (1): в линейной регрессии  $Y$  должен быть линейно связан с  $X$ . Условие (Б) относится к статистическим свойствам коэффициентов  $b_0$  и  $b_1$  модели (1). В настоящей работе речь идет только о проверке выполнения условия (А).

Ситуации на рис. 1а и 1б по-разному могут быть сопоставлены с условием применимости (А). В ситуации рис. 1а линейность связи может быть оценена визуально (экспертно) [9, 18]; статистическая значимость отклонения от линейности может быть проверена, например, с помощью корреляционного отношения [19]. Ситуация, изображенная на рис. 1б, не позволяет визуально (экспертно) оценить линейность связи  $Y$  и  $X$  в той мере, в какой это возможно для ситуации рис. 1а.

В принципе, и здесь возможно применение корреляционного отношения; отметим лишь, что идеология корреляционного отношения предполагает разделение значений предиктора  $X$  на интервалы (страты), в каждом из которых рассчитываются средние значения  $Y$  и  $X$ . После такой стратификации данные рис. 1б становятся похожими на данные рис. 1а. В отличие от данных рис. 1а, где разделение на страты происходит в процессе планирования эксперимента, стратификация данных рис. 1б – субъективный (неоднозначный) выбор биостатистика. Отметим, что экспертная (визуальная) оценка специалистом каких-либо статистических гипотез не является чем-то второстепенным (по сравнению со статистической проверкой гипотез), что подтверждается рекомендациями известных статистиков [9, 18].

Кроме того, в известных статистических пакетах (Statistica, SPSS) не было найдено аналогов проверки линейности связи  $Y$  с  $X$  для данных типа рис. 1б через процедуру расчета корреляционного отношения с последующим применением скользящего среднего [20–22]. Анализ научных публикаций последнего десятилетия (2012–2022 гг.) показал лишь небольшое количество работ, содержащих описание линейной регрессии, в которых авторы сообщали о проверке линейности связи

Y и X с помощью корреляционного отношения. Запрос на сервисе полнотекстового поиска по научной литературе Google Scholar, включающий термины (и их синонимы, приведенные в скобках): linear regression и linearity test (test of linearity, linearity check, check of linearity, linearity validation, validation of linearity, linearity verification, verification of linearity), показал в результирующей выдаче от 14 до 7 960 документов, в зависимости от формулировки. При добавлении в поисковый запрос термина correlation ratio количество документов варьировалось от 0 до 10. Всего было найдено 13 статей, включавших все три понятия (linear regression, linearity test, correlation ratio), и только в 5 из них корреляционные отношения использовались для проверки линейности связи. Более того, во многих публикациях не проводилась даже визуальная (экспертная) оценка линейности связи Y и X с помощью диаграмм типа рис. 1б [23–25]; при этом в публикациях [23, 25] приводятся и анализируются значения коэффициентов линейной регрессионной модели (или коэффициентов корреляции Пирсона между Y и X), которые по определению имеют смысл только при наличии линейной связи типа (1). Между тем в некоторых случаях простейший анализ диаграммы рассеяния может показать нелинейность связи Y и X и необходимость преобразования переменных для использования линейной регрессии [26].

Внешняя схожесть рис. 1а и 1б не должна вводить в заблуждение: на рис. 1а значения X осознанно выбраны для данного эксперимента и отражают объективную реальность. При разделении на страты данных рис. 1б количество страт и их границы – это субъективный выбор биостатистика, осуществляющего анализ данных; очевидно, что различные способы стратификации будут давать различные результаты при проверке линейности связи Y и X.

Тем не менее математические методы построения линейных регрессионных моделей типа (1) в обоих случаях, изображенных на рис. 1а и рис. 1б, одинаковы (метод наименьших квадратов, метод наибольшего правдоподобия). Для конкретных экспериментальных данных, состоящих из пар значений  $(x_i, y_i)$ , модель линейной регрессии записывается в виде [8]:

$$y_i = b_0 + b_1 x_i + \varepsilon_i, \quad (2)$$

где  $x_i$  – значение предиктора X для  $i$ -го объекта исследования,  
 $y_i$  – соответствующее значение отклика Y,  
 $b_0$  и  $b_1$  – коэффициенты модели линейной регрессии,  
 $\varepsilon_i$  – ошибки интерполяции (модели).

Используя процедуры нахождения коэффициентов модели (2) – метод наименьших квадратов, или метод максимального правдоподобия, получаем расчетные  $\hat{y}_i$  значения отклика Y в виде:

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 \cdot x_i.$$

В рамках линейных моделей (1), (2) изменение Y при изменении X описывается одним параметром  $\hat{b}_1$  для любого X (приращение Y при изменении X на единицу равно  $\hat{b}_1$  вне зависимости от значения X). При любых других формах связи Y и X, отличных от (1), (2), изменение Y при изменении X оказывается разным для разных X.

Достаточным (но не обязательно необходимым) условием линейности Y по X является двумерная нормальность распределения значений Y и X. Доказать это на реальных биологических или медицинских данных строго не представляется воз-

возможным. А. Afifi с соавторами [9] предлагают визуальный способ оценки двумерной нормальности – на диаграмме рассеяния в координатах  $X$ – $Y$  (как на рис. 1б) точки должны заполнять эллипс. Иногда довольствуются проверкой нормальности распределения отдельно  $Y$  и  $X$  [7]. Тем не менее проблема установления (проверки) линейности связи  $Y$  и  $X$  остается актуальной.

В настоящей работе для проверки условия (А) линейности связи  $Y$  и  $X$  для данных типа рис. 1б предлагается использовать методы скользящего среднего. Как будет показано, для данных регрессии типа рис. 1б скользящее среднее позволяет выявить линейность связи  $Y$  и  $X$  с применением минимального количества субъективных факторов.

### Скользящее среднее для независимых наблюдений

Методы скользящего среднего широко применяются в регрессионном анализе [17, 27]. Наибольшее применение методы скользящего среднего получили при анализе временных рядов, где они используются для сглаживания случайных колебаний изучаемого фактора и выявления трендов [17, 18]. В данной работе методы скользящего среднего применяются для данных независимых наблюдений (не временных рядов) и используются для решения «обратной» (по сравнению с временными рядами) задачи: задачи выявления «структуры» связи между переменными, которая в первичных данных обычно скрыта большой дисперсией обеих переменных  $X$  и  $Y$ . Скользящее среднее позволяет уменьшить дисперсию переменных, в результате чего возможная связь между  $Y$  and  $X$  становится более явной.

Рассмотрим процедуру построения скользящего среднего для первичных данных, представляющих независимые наблюдения (не временные ряды). Пусть имеются пары значений  $(x_i, y_i)$  для  $n$ -наблюдений переменных  $X$  и  $Y$  ( $i = 1, 2, 3, \dots, n$ ). Для построения скользящего среднего первичные значения предиктора  $X$  ( $x_1, x_2, \dots, x_n$ ) упорядочиваются по возрастанию, после чего область изменения  $X$  разделяется на страты; в каждой страте рассчитываются средние значения  $Y$  и  $X$ .

В первой страте средние значения предиктора  $\langle X(1) \rangle$  и отклика  $\langle Y(1) \rangle$  определяются формулами:

$$\langle X(1) \rangle = \frac{1}{n_w} \sum_{i=1}^{n_w} x_i, \quad \langle Y(1) \rangle = \frac{1}{n_w} \sum_{i=1}^{n_w} y_i, \quad (3)$$

где индекс «1» у средних значений означает номер страты;

$n_w$  – размер «окна» усреднения (window).

Таким образом, в первую страту попадают  $n_w$  объектов с минимальными значениями предиктора  $X$ . Во второй страте средние значения  $\langle X(2) \rangle$  и  $\langle Y(2) \rangle$  определяются формулами:

$$\langle X(2) \rangle = \frac{1}{n_w} \sum_{i=n_s+1}^{n_w+n_s} x_i, \quad \langle Y(2) \rangle = \frac{1}{n_w} \sum_{i=n_s+1}^{n_w+n_s} y_i, \quad (4)$$

где  $n_s$  – шаг (step) при построении скользящего среднего.

Таким образом, вторая страта получается путем сдвига начала отсчета на величину  $n_s$ . Средние значения остальных страт определяются аналогично при сдвиге на  $n_s$  для каждой новой страты. В результате первичные значения  $X$  и  $Y$  ( $x_1, x_2, \dots, x_n; y_1, y_2, \dots, y_n$ , всего  $n$ -значений) заменяются на набор средних значений  $\langle X(N_s) \rangle$

и  $\langle Y(N_s) \rangle$  показателей  $X$  и  $Y$  в стратах ( $N_s$  – номер страты). При этом разброс индивидуальных значений переменных  $X$  и  $Y$  уменьшается и характер связи между ними становится более явным.

### Применение скользящего среднего к данным линейной регрессии

В качестве примера применения скользящего среднего вернемся к данным рис. 1б (статистическая связь веса и роста). На рис. 2а и рис. 2б показаны данные скользящего среднего для показателя «вес» по предиктору «рост» с размером окна  $n_W = 25$  наблюдений и с шагом скользящего среднего, равного  $n_S = 25$  (получается 31 среднее значение веса и роста) и  $n_S = 1$  (351 среднее значение). Для расчета скользящих средних в среде Statistica for Windows использована компьютерная программа [28].

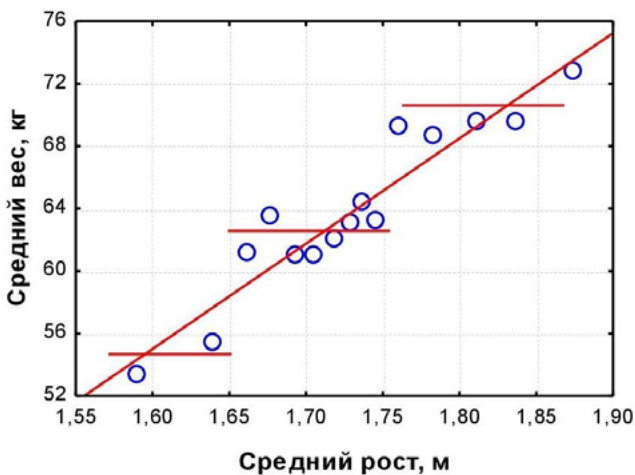


Рис. 2а. Соотношение средних значений веса и роста подростков. Окно усреднения равно  $n_W = 25$ , шаг  $n_S = 25$

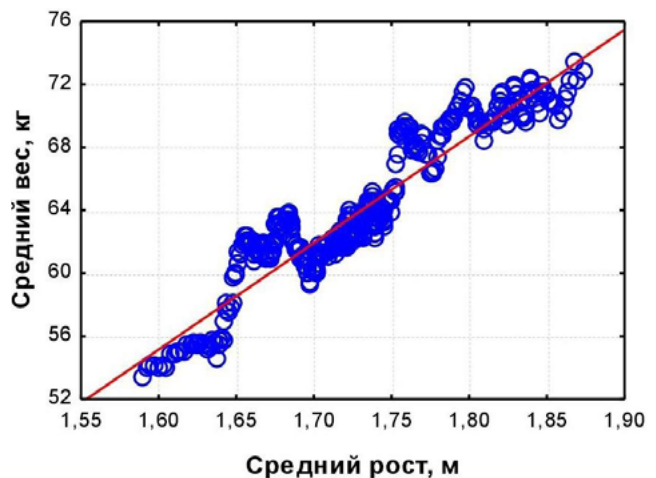


Рис. 2б. Соотношение средних значений веса и роста подростков. Окно усреднения равно  $n_W = 25$ , шаг  $n_S = 1$

На рис. 2а значения предиктора  $X$  (рост юношей из рис. 1б) разделены на непересекающиеся страты размером 25 наблюдений, в каждой из которых рассчитаны средние значения роста и веса. Рис. 2а в какой-то мере похож на рис. 1а (для одного значения роста имеем одно среднее значение веса), хотя значения роста не расположены здесь так регулярно, как на рис. 1а. Рис. 2б наглядно показывает, что точной линейной связи средних значений веса и роста нет; более того, на рис. 2б можно видеть три области средних значений роста (это области значений роста до 165 см, затем от 165 до 175 см, и наконец область выше 175 см), в каждой из которых средние значения веса слабо меняются при изменении роста, при этом имеются значительные различия веса между областями. Таким образом, рис. 2б показывает более подробную (чем рис. 2а) картину связи средних значений роста и веса. Среди 351 точки на рис. 2б есть, очевидно, та 31 точка, которая показана на рис. 2а.

Скользящее среднее (рис. 2а) дает более детальную картину связи  $X$  со средними значениями  $Y$  по сравнению с первичными данными (рис. 1б), но уравнения регрессии в обоих случаях похожи (что и следовало ожидать): связь веса и роста на рис. 1б описывается соотношением:

$$\text{вес(кг)} = -50,6 + 0,661 \cdot \text{рост(см)},$$

а на рис. 2а – соотношением:



$$\text{вес(кг)} = -53,0 + 0,675 \cdot \text{рост(см)}.$$

Результаты применения процедуры скользящего среднего к реальным биологическим или медицинским данным зависят от выбора размера «окна» усреднения  $n_W$  и от выбора шага  $n_S$ . Что касается выбора окна  $n_W$ , мы предложили подход с использованием кумулятивной функции для  $Y$  по  $X$  (подробности см. в [29]). Шаг усреднения  $n_S$  можно в любом случае выбирать равным единице, как на рис. 2б (получаем максимум информации о средних значениях  $Y$  и  $X$ ). Для «укрупненного» анализа размер шага скользящего среднего  $n_S$  можно увеличить, как показано на рис. 2а. В результате такого выбора  $n_W$  и  $n_S$  неоднозначность результатов скользящего среднего практически устраняется.

### 3. Логистическая регрессия

Одной из наиболее распространенных моделей нелинейной регрессии является логистическая регрессия. Логистическая регрессия применяется для описания статистической связи между количественным предиктором  $X$  и дихотомическим откликом  $Y$  ( $Y$  принимает два значения:  $Y = 0$  и  $Y = 1$ ). Данные такого рода часто встречаются в эпидемиологических исследованиях, когда дихотомический  $Y$  кодирует наличие или отсутствие некоторого заболевания, а  $X$  – фактор риска возникновения заболевания. Обычно считается, что  $Y = 1$  кодирует наличие заболевания, а  $Y = 0$  – его отсутствие у конкретного пациента.

В модели логистической регрессии статистическая связь между  $Y$  и  $X$  предполагается в виде:

$$W(Y=1|X) = \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)}, \quad (5)$$

где  $W(Y=1|X)$  – вероятность обнаружения в экспериментальных данных значения  $Y = 1$  для заданного значения  $X$  [9, 30].

При использовании соотношения (5) имеем

$$\log\left(\frac{W}{1-W}\right) = b_0 + b_1 \cdot X. \quad (6)$$

Таким образом, как и в случае линейной регрессии, имеем линейную связь, но не между  $X$  и  $Y$ , а между предиктором  $X$  и комплексом  $\log\left(\frac{W}{1-W}\right)$ , который называют логитом ( $Y$ ), это условие применимости логистической регрессии. Каких-либо ограничений на тип предиктора  $X$  обычно не предъявляется [30]; напомним, что в линейной регрессии для получения нужных статистических свойств коэффициентов регрессии требуется нормальность распределения  $Y$  и  $X$ .

Известно достаточное (но не обязательно необходимое) условие, при котором выполняется линейное соотношение (6): распределение  $X$  в группах  $Y = 0$  и  $Y = 1$  подчиняется нормальному закону с равными дисперсиями [31]. В остальных случаях линейность связи  $X$  с логитом ( $Y$ ) необходимо проверять.

Аналогично случаю линейной регрессии при выполнении условия линейности логита изменение логита ( $Y$ ) при изменении  $X$  описывается одним параметром, оди-

наковым для любого  $X$ . Это параметр – отношение шансов OR, который для изменения  $X$  на единицу рассчитывается по формуле:

$$OR = \exp(b_1),$$

где  $b_1$  – коэффициент модели (6).

На рис. 3а представлены типичные первичные данные для логистической регрессии (количественный  $X$  и дихотомический  $Y$ ), взятые из монографии [30]. Здесь отклик  $Y$  представляет заболевание сердечно-сосудистой системы (ССС) у пациентов данного возраста:  $Y = 0$  означает отсутствие заболевания,  $Y = 1$  – наличие заболевания у данного пациента. Очевидно, что данные типа рис. 3а не позволяют оценить выполнение соотношения (6) даже в той минимальной мере, в какой это возможно для линейной регрессии для данных рис. 1б. Действительно, наличие у отклика  $Y$  всего двух значений (нуль и единица) исключает всякую возможность экспертной (визуальной) оценки формы связи  $Y$  и  $X$ . Кривая логистической регрессии на рис. 3а проведена согласно модели (5), однако нет никакой уверенности, что вероятность иметь заболевание ССС описывается именно такой кривой, если опираться на визуальный анализ нулей и единиц.

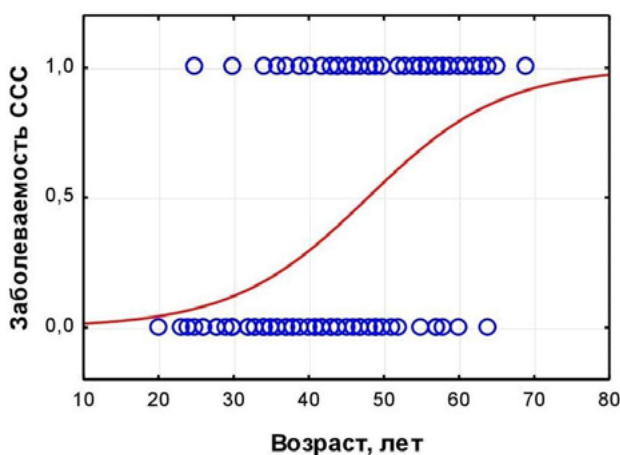


Рис. 3а. Распределение заболеваемости ССС в зависимости от возраста пациента [30]. Сплошная линия – линия логистической регрессии

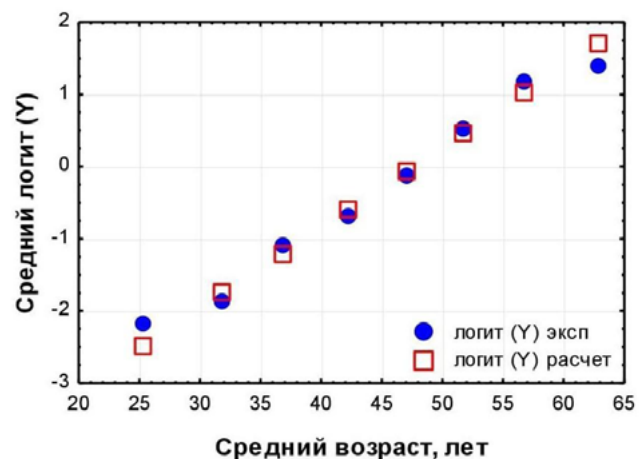


Рис. 3б. Стратификация 100 пациентов на 8 страт по возрастным группам [30]. Синие кружки – средние значения логита заболеваемости ССС и возраста в стратах. Красные квадраты – расчеты методом логистической регрессии

В отличие от линейной регрессии, в логистической регрессии уделяется внимание проверке линейности связи логита ( $Y$ ) с  $X$ , т. е. выполнению соотношения (6). Многие известные статистические пакеты (SAS, SPSS) включают опции, аналогичные процедуре корреляционного отношения в линейной регрессии [20]. Эти опции предполагают деление предиктора  $X$  на непересекающиеся страты (аналогично тому, как это делается при расчете корреляционного отношения в линейной регрессии) и расчет статистических критериев согласия для проверки гипотезы о линейности связи логита ( $Y$ ) с  $X$  (критерий хи-квадрат Хосмера – Лемешова [30], а также индексы  $R^2$  [32, 33]. Пользователи логистической регрессии активно пользуются критериями согласия, однако ограничиваются лишь определением значений критерия согласия и уровня значимости  $p$  без графического анализа полученных результатов (иногда данные стратификации представляют в виде таблицы средних значе-

ний  $X$  и логита ( $Y$ ) в сравнении с результатами логистической регрессии, но без графического представления).

На рис. 3б показано сравнение данных логистической регрессии с данными стратификации предиктора  $X$ . Предиктор  $X$  (возраст пациента) в работе [30] был разделен на 8 страт, в каждой страте рассчитаны средние значения возраста и логита ( $Y$ ), которые являются результатом усреднения первичных экспериментальных данных. На рис. 3б приведены также значения логита ( $Y$ ), вычисленные в этих же стратах методом логистической регрессии. В результате перехода от вероятности  $W$  (как на рис. 3а) к логиту ( $Y$ ) на рис. 3б логистическая кривая превращается в прямую. Согласно визуальной оценке экспериментальные данные для логита ( $Y$ ) в целом показывают линейный рост с ростом возраста. Статистическая проверка гипотезы о линейности связи логита ( $Y$ ) с возрастом это подтверждает: согласно критерию согласия хи-квадрат Хосмера – Лемешова  $\chi^2 = 0,16$  при 6 степенях свободы, при котором гипотеза о линейности не отвергается на обычном уровне значимости  $\alpha = 0,05$ ,  $p > 0,05$ .

На примере данных Хосмера – Лемешова (рис. 3б) можно продемонстрировать неоднозначность результатов проверки гипотезы о линейности логита, которые получаются при различных способах стратификации данных. Хосмер и Лемешов разделили пациентов на 8 непересекающихся страт по возрастным категориям (20–29 лет, 30–34 г. и т. д.), каждая страта содержит от 8 до 17 пациентов, т. е. неравное число наблюдений в страте. Если обратиться к известным статистическим пакетам, то в них предиктор  $X$  обычно разделяется на страты с равным числом наблюдений [34]. В примере Хосмера – Лемешова общее число пациентов равно 100, так что в каждой из 8 страт должно быть 12 или 13 пациентов. Результаты деления  $X$  на 8 страт с равным числом пациентов представлены на рис. 4а; в некоторых точках заметно явное отличие от рис. 3б. Мы провели расчеты критерия согласия Хосмера – Лемешова для 8 страт с равным числом пациентов в них и обнаружили резкое увеличение критерия  $\chi^2 = 1,87$  (это значение критерия  $\chi^2$  в 11 раз больше, чем у Хосмера – Лемешова). При таком значении  $\chi^2$  нулевая гипотеза о линейной связи логита ( $Y$ ) с возрастом также не отвергается (из-за недостаточного числа пациентов), но вероятность отвергнуть нулевую гипотезу значительно повышается.

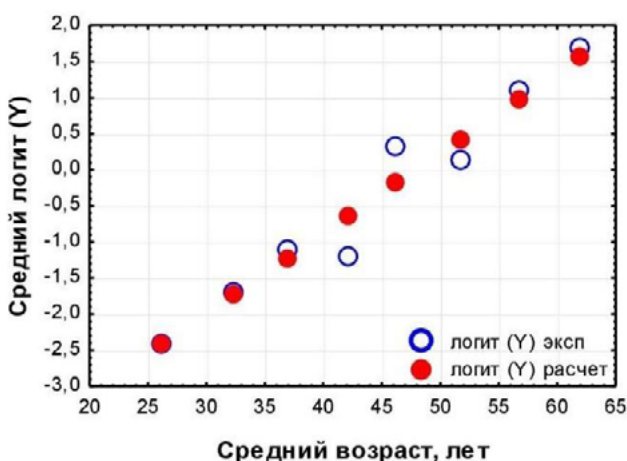


Рис. 4а. Стратификация 100 пациентов на 8 страт, равных по численности пациентов. Обозначения как на рис. 3б

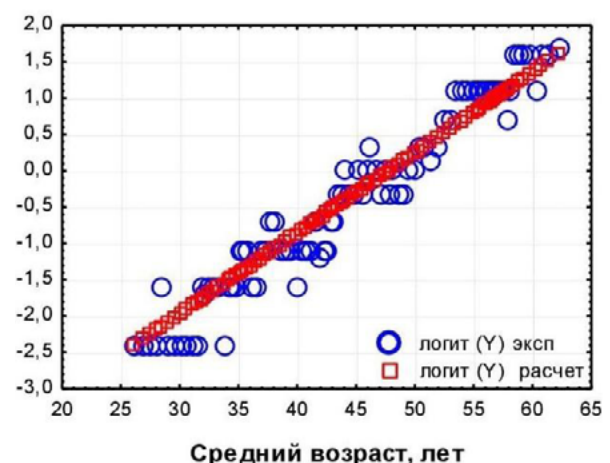


Рис. 4б. Средний логит вероятности иметь заболевание ССС по результатам скользящего среднего (окно  $n_W = 12$ , синие кружки). Красные квадраты – расчеты методом логистической регрессии

Поскольку экспериментальные данные прошли проверку на линейность связи логита с  $X$ , мы имеем право построить модель типа (6), описывающую связь логита вероятности  $W$  иметь заболевание ССС с возрастом:

$$\log\left(\frac{W}{1-W}\right) = b_0 + b_1 \cdot X = -5,309 + 0,1109 \cdot \text{возраст}. \quad (7)$$

### Применение скользящего среднего к данным логистической регрессии

Применение скользящего среднего к данным логистической регрессии имеет особенность по сравнению с линейной регрессией – процедура скользящего среднего переводит дихотомическую переменную  $Y$  в количественную переменную. Рассмотрим несколько примеров применения скользящего среднего к реальным эпидемиологическим данным; каждый пример демонстрирует особенности данных и способы их выявления методами скользящего среднего.

**Пример 1.** Применим процедуру скользящего среднего к данным Хосмера – Лемешова. На рис. 4б показаны результаты скользящего среднего для отклика  $Y$  = заболевание ССС по предиктору «возраст» вместе с результатами логистической регрессии. Среди 89 точек на этом рисунке имеются 8 точек, которые есть на рис. 3б; кроме этого, здесь есть еще 81 точка для промежуточных значений возраста.

Данные рис. 4б можно использовать для построения модели связи логита вероятности заболевания ССС с возрастом. Из данных рис. 4б (синие точки скользящего среднего) получаем:

$$\log\left(\frac{W}{1-W}\right) = b_0 + b_1 \cdot X = -5,632 + 0,1179 \cdot \text{возраст}. \quad (8)$$

Очевидно, соотношение (8), полученное для данных скользящего среднего, практически не отличается от соотношения (7), полученного методом логистической регрессии. Этого следовало ожидать, поскольку график рис. 4б подтверждает справедливость соотношений типа (5), (6), которые являются условиями применимости логистической регрессии.

**Пример 2.** В предыдущем примере результаты стратификации в духе логистической регрессии хорошо согласуются с результатами скользящего среднего. На данном примере мы покажем отличия результатов скользящего среднего от результатов регрессии, притом что линейность логита в экспериментальных данных подтверждается критериями согласия.

В 2002 г. J. Peng с соавторами в журнале обучающей направленности опубликовали методическую статью [34] по применению логистической регрессии с подробным описанием тех процедур, которые необходимо выполнить при решении конкретной задачи, и тех показателей, которые необходимо отразить в сводке результатов. В частности, на примере конкретных эпидемиологических данных в этой работе была построена модель логистической регрессии для связи некоторого дихотомического отклика  $Y$  (*remedia*) с количественным предиктором  $X$  (*reading score*):

$$\log\left(\frac{W}{1-W}\right) = b_0 + b_1 \cdot X = -0,997 + 0,0281 \cdot \text{reading score}, \quad (9)$$

коэффициент регрессии при  $X$  оказался статистически значим ( $p = 0,0245$  по  $t$ -критерию).

Были также рассчитаны критерии согласия для проверки линейности связи предиктора  $X$  с логитом ( $Y$ ). Все критерии показали, что при разделении предиктора  $X$  на 10 групп (страт) с примерно одинаковым числом наблюдений гипотеза о линейности логита ( $Y$ ) не отвергается.

В статье [34] не приведены графики связи средних значений предиктора  $X$  с логитом ( $Y$ ), но они оказались достаточно показательными при их построении. На рис. 5а показаны результаты, полученные нами для этих данных после такой же стратификации, как и в оригинальной статье [34]. Видно, что логит ( $Y$ ) при увеличении  $X$ , конечно, имеет тенденцию к уменьшению, но явной линейности здесь не наблюдается. Кроме того, расчетные данные во многих стратах весьма далеки от эксперимента (велик разброс экспериментальных значений около линии логистической регрессии).

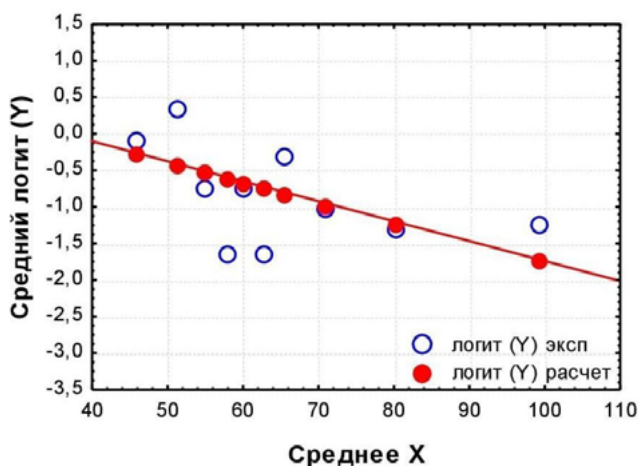


Рис. 5а. Красные кружки – расчетные данные логистической регрессии, синие кружки – результат стратификации на 10 страт [34]

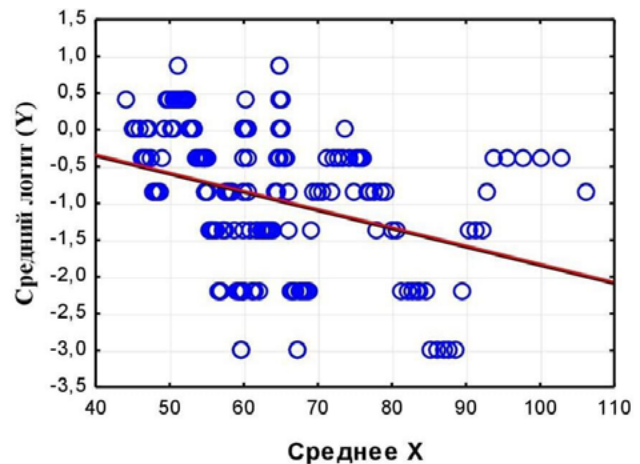


Рис. 5б. Кружки – результат скользящего среднего (эксперимент), сплошная линия – линейная интерполяция

Применим к данным [34] процедуру скользящего среднего, результат показан на рис. 5б. Этот график заставляет еще сильнее усомниться в строгой линейности связи между логитом ( $Y$ ) и количественным предиктором  $X$ : на рис. 5б видны явные максимумы средних значений логита ( $Y$ ) при  $X$ , равных 50, 62, 74 и 100, и минимумы логита при  $X$ , равных 60, 68 и 85–90. Особенно четко нелинейность логита ( $Y$ ) видна для значений  $X$ , превышающих 65. Связь экспериментальных значений логита ( $Y$ ) с  $X$  на рис. 5б в чем-то похожа на таковую на рис. 5а, но рис. 5б более подробный. В частности, размах значений логита ( $Y$ ) на рис. 5б (от  $-3$  до  $+1$ ) гораздо больший, чем на рис. 5а (от  $-1,7$  до  $+0,4$ ).

Кажется, что данные логистической регрессии (сплошная линия на рис. 5а) кардинально отличаются от данных скользящего среднего (рис. 5б). Если, однако, аппроксимировать данные рис. 5б прямой линией, получается следующее соотношение:

$$\log\left(\frac{W}{1-W}\right) = b_0 + b_1 \cdot X = -0,631 + 0,0248 \cdot \text{reading score}, \quad (10)$$

где значение коэффициента при  $X$ , равное  $b_1 = -0,0248$ , оказывается близко к соответствующему значению  $b_1 = -0,0281$  в логистической регрессии для первичных данных в формуле (9). Таким образом, данные скользящего среднего (10) согласу-

ются с данными логистической регрессии (9), но скользящее среднее позволяет выявить более детально «структуру» связи логита ( $Y$ ) с  $X$ .

**Пример 3.** В предыдущем примере показано, что графическое представление данных в логистической регрессии позволяет оценить возможную нелинейность связи логита ( $Y$ ) с предиктором  $X$  даже в случае подтверждения линейности с помощью критериев согласия. В данном примере показано, что даже при выполнении всех условий применимости логистической регрессии результаты логистической регрессии могут *кардинально* отличаться от первичных данных. Это отличие удается установить только благодаря процедуре скользящего среднего. При отсутствии анализа скользящего среднего логистическая регрессия дает некорректный результат, который не согласуется с реальными данными.

Для построения модели использован эпидемиологический материал, включающий 100 женщин в возрасте старше 60 лет, для которых определялись антропометрические показатели и распространенность различных соматических патологий [35]. В данном примере описывается связь распространенности заболеваний щитовидной железы (ЩЖ) с индексом массы тела (ИМТ) женщин. Заболевания ЩЖ кодируются символом 0, если заболевания у данной женщины нет, и 1, если заболевание есть.

Нами был рассчитан критерий согласия Хосмера – Лемешова для проверки линейности связи предиктора  $X$  (ИМТ) с логитом вероятности иметь заболевание ЩЖ, который показал, что при разделении предиктора ИМТ на 10 страт с одинаковым числом наблюдений гипотеза о линейности логита не отвергается.

Применение метода логистической регрессии дает следующие результаты. Соотношение типа (6) для данной задачи имеет вид ( $W$  – вероятность иметь заболевание ЩЖ):

$$\log\left(\frac{W}{1-W}\right) = b_0 + b_1 \cdot X = -0,319 + 0,0111 \cdot \text{ИМТ}. \quad (11)$$

Коэффициент регрессии  $b_1$  статистически значимо не отличается от нуля,  $p = 0,877$ . Таким образом, согласно результатам логистической регрессии не существует значимой статистической связи между распространенностью заболеваний щитовидной железы с индексом массы тела женщин.

Сказанное позволяет сделать вывод: метод логистической регрессии использован при выполнении условий применимости, но значимой связи между  $W$ (ЩЖ) и ИМТ не обнаружено.

Проведем графический анализ связи  $W$ (ЩЖ) с ИМТ. На рис. 6а показаны результаты стратификации данных на 10 групп по 10 пациентов в каждой, а также расчетные данные методом логистической регрессии. Видно, что экспериментальные данные во многих случаях весьма далеки от расчетных данных.

Теперь применим к первичным данным процедуру скользящего среднего по ИМТ. На рис. 6б представлена связь логита вероятности  $W$ (ЩЖ) с ИМТ после выполнения операции скользящего среднего. Экспертная оценка данных рис. 6б показывает, что связь логита  $W$ (ЩЖ) иметь заболевание щитовидной железы с ИМТ у женщин существует, однако она не только не является линейной (как того требует логистическая регрессия), но даже не является монотонной: в интервале значений ИМТ от 22 до 26 кг/м<sup>2</sup> логит  $W$ (ЩЖ) в среднем уменьшается при увеличении ИМТ, а в интервале ИМТ от 26 до 30 кг/м<sup>2</sup> логит  $W$ (ЩЖ) возрастает.



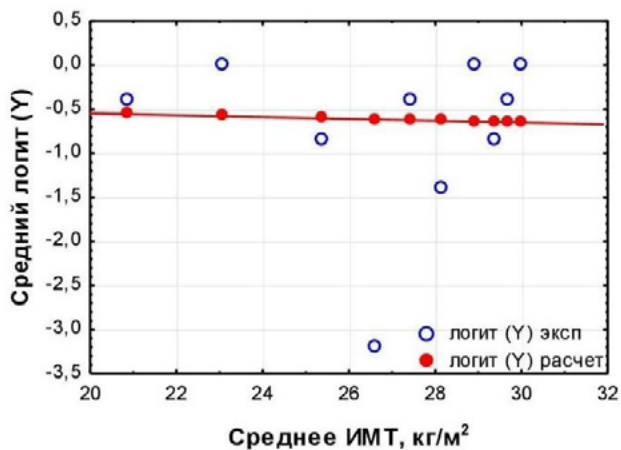


Рис. 6а. Связь логита  $W$ (ЩЖ) с ИМТ для данных 10 страт. Синие кружки (эксперимент) – средние значения логита  $W$ (ЩЖ), красные кружки – расчет методом логистической регрессии

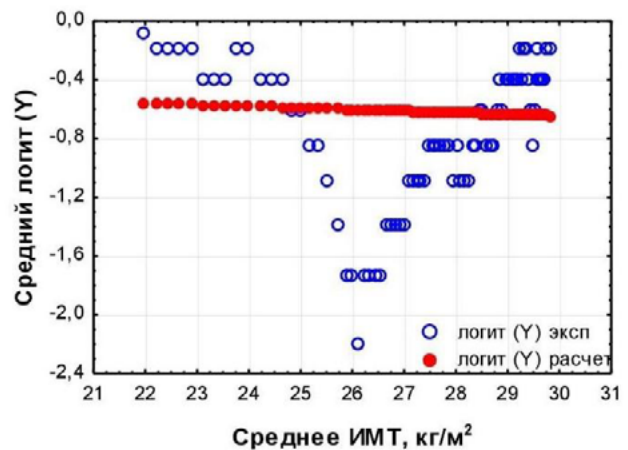


Рис. 6б. Связь логита  $W$ (ЩЖ) с ИМТ для данных 81 страты скользящего среднего. Синие кружки (эксперимент) – средние значения логита  $W$ (ЩЖ), красные кружки – расчет методом логистической регрессии

Проведем анализ статистической значимости различий  $W$ (ЩЖ) по данным скользящего среднего для различных значений ИМТ. Возьмем первую страту по ИМТ на рис. 6б. В этой страте среднее значение ИМТ = 21,90 кг/м<sup>2</sup>, распространенность патологии ЩЖ в этой страте максимальна и равна  $W = 0,476$ . Возьмем 25-ю страту по ИМТ, среднее значение ИМТ = 26,10 кг/м<sup>2</sup>, распространенность патологии ЩЖ в этой страте минимальна и равна  $W = 0,143$ . Страты № 1 и 25 содержат по 20 наблюдений и не пересекаются. По двустороннему критерию сравнения долей  $W = 0,476$  и  $W = 0,143$  различаются статистически значимо,  $p = 0,0195$ . В последней страте № 81 (среднее значение ИМТ = 29,83 кг/м<sup>2</sup>) распространенность патологии ЩЖ равна  $W = 0,429$ . Различие со стратой № 25 статистически значимо,  $p = 0,0404$ .

Сравним выводы, получаемые для данного эпидемиологического материала методом логистической регрессии и методами скользящего среднего. Применение логистической регрессии показывает, что экспериментальные данные согласуются с предположением о линейности логита, но получаемый коэффициент регрессии не отличается от нуля статистически значимо. Это означает, что статистически значимой связи  $W$ (ЩЖ) и ИМТ нет. Применение методов скользящего среднего дает противоположный результат: логит  $W$ (ЩЖ) является не линейной, а V-образной функцией ИМТ; имеет место статистически и предметно значимая связь между распространенностью заболеваний щитовидной железы и ИМТ у пожилых женщин. Таким образом, основываясь только на результатах логистического регрессионного анализа, можно было «не заметить» явную связь между  $W$ (ЩЖ) и ИМТ.

#### 4. Заключение

Линейная и нелинейная (логистическая) регрессия, описывающая статистическую связь отклика  $Y$  с предиктором  $X$ , должна использоваться в области своей применимости. Основное условие применимости данных регрессионных моделей – линейность связи  $Y$  с  $X$  в линейной регрессии и линейность связи логита ( $Y$ ) с  $X$  в логистической регрессии. При обнаружении линейной связи между  $Y$  и  $X$  изменение отклика при изменении предиктора может быть описано одним коэффициентом (одним числом) для всей области изменения предиктора  $X$ . Если же связь нелинейная,

изменение отклика описывается разными значениями при различных значениях  $X$ . Поэтому так важно знать, является ли связь между  $Y$  и  $X$  линейной или нет. Если показатели линейной связи применяются к нелинейным данным, выводы статистического анализа будут некорректны.

Процедура проверки линейности зачастую является обособленным этапом анализа данных, видимо, поэтому во многих публикациях авторы вообще не сообщают о результатах проверки линейности связи  $X$  и  $Y$ , если таковая проверка вообще проводилась (например, работы [23, 25]).

В этой ситуации метод скользящего среднего может показать характер связи  $Y$  и  $X$  более наглядно и явно, чем обычная диаграмма рассеяния (сравнение рис. 1б с рис. 2а). Вместе с тем даже элементарный анализ диаграммы рассеяния может показать явную нелинейность связи  $Y$  и  $X$  (например, [26]).

Что касается методов проверки линейности связи для логистической регрессии, такая проверка заложена во многих статистических пакетах, для чего используются различные статистические критерии: критерий хи-квадрат Хосмера – Лемешова [30], индексы  $R^2$  [32, 33]. При этом во многих публикациях не сообщается о проверке линейности в логистической регрессии (например, [36, 37]).

В данной работе показано, что ориентация только на статистические критерии может привести к искажению выводов логистической регрессии. В частности, критерии согласия могут не отвергать гипотезу о линейности связи предиктора  $X$  с логитом ( $Y$ ), а графический анализ показывает явно нелинейную связь (пример 3).

Активными пропагандистами графического представления и анализа результатов статистического анализа были такие известные специалисты, как Дж. Тьюки и Э. Сигел. Американский статистик, профессор нескольких университетов Э. Сигел в монографии [18] практически каждый вывод иллюстрирует графиками с подробными комментариями. Один из создателей современной науки анализа данных Дж. Тьюки в монографии [38] писал: «Графики, подчеркивающие лишь то, что нам уже известно, нередко не стоят места, которое они занимают. Графики, которые надо рассматривать с лупой, заставляют нас тратить понапрасну время и мало полезны. График имеет наибольшую ценность тогда, когда он вынуждает нас заметить то, что мы совсем не ожидали увидеть». Приведенные выше рис. 2, 4, 5 и особенно 6 как раз и показывают то, что мы не ожидали увидеть.

## 5. Список литературы

1. Osborne, J. W. Multiple Regression Assumptions / J. W. Osborne, E. Waters. – ERIC Digest, 2002.
2. Park, C. C. A linearity test statistic in a simple linear regression / C. C. Park, K. E. Lee // J. of the Korean Data and Information Science Society. – 2014. – Vol. 25, No. 2. – P. 305–315. – DOI 10.7465/jkdi.2014.25.2.305.
3. Nimon, K. F. Statistical assumptions of substantive analyses across the general linear model: a mini-review / K. F. Nimon // Front. Psychology. – 2012. – Vol. 3, No. 322. – DOI 10.3389/fpsyg.2012.00322.
4. Dhakal, C. P. Regression invented as statistics / C. P. Dhakal // International J. of Interdisciplinary Research and Innovations. – 2018. – Vol. 6, No. 2. – P. 1–5.
5. Stanton, J. M. Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors / J. M. Stanton // J. of Statistics Education. – 2001. – Vol. 9, No. 3. – DOI 10.1080/10691898.2001.11910537.

6. Aldrich, J. Fisher and regression / J. Aldrich // *Statistical Science*. – 2005. – Vol. 20, No. 4. – P. 401–417.
7. Draper, N. R. Applied Regression Analysis, 3<sup>rd</sup> ed. / N. R. Draper, H. Smith. – Wiley-Interscience Publication, 1998. – 736 p. – ISBN 978-0-471-17082-2.
8. Afifi, A. A. Statistical Analysis: A Computer Oriented Approach / A. A. Afifi, S. P. Azen. – New York : Academic Press, 1972. – 366 p. – ISBN: 9780120444502.
9. Afifi, A. A. Computer-aided multivariate analysis, 4<sup>th</sup> ed. / A. A. Afifi, S. May, V. Clark. – Chapman&Hall/CRC, 2003. – 512 p. – ISBN 978-1584883081.
10. An Analysis of the Links between Smoking and BMI in Adolescents: A Moving Average Approach to Establishing the Statistical Relationship between Quantitative and Dichotomous Variables / A. N. Varaksin, E. D. Konstantinova, T. A. Maslakova [et al.] // *Children (Basel)*. – 2022. – Vol. 9, No. 2. – DOI 10.3390/children9020220. – EDN GKBQJD.
11. Владимиров, В. Г. Радиозащитные рецептуры. Оптимизация состава и механизм действия / В. Г. Владимиров, Г. А. Поддубский, Г. И. Разоренов. – Л. : Воениздат, 1988. – 144 с.
12. Налимов, В. В. Теория эксперимента / В. В. Налимов. – М. : Наука, 1971. – 208 с.
13. Are in vivo and in vitro assessments of comparative and combined toxicity of the same metallic nanoparticles compatible, or contradictory, or both? A juxtaposition of data obtained in respective experiments with NiO and Mn<sub>3</sub>O<sub>4</sub> nanoparticles / I. A. Minigalieva, T. V. Bushueva, E. Fröhlich [et al.] // *Food and Chemical Toxicology*. – 2017. – Vol. 109, No. 1. – P. 393–404. – DOI 10.1016/j.fct.2017.09.032. – EDN ZGWVNH.
14. Fletcher, R. H. Clinical epidemiology: The essentials / R. H. Fletcher, S. W. Fletcher. – 5<sup>th</sup> ed. – Lippincott Williams & Wilkins, 2012. – 272 p. – ISBN 978-1451144475.
15. Расина, Л. Н. Статистический анализ эколого-физиологических данных в интерпретации техногенных воздействий / Л. Н. Расина, Н. А. Орехова, А. Н. Варакин // *Экологические системы и приборы*. – 2011. – № 12. – С. 50–53. – EDN SHXDFX.
16. Effects of Environmental Radioactive Pollution on the Cardiovascular Systems of Ural Region Residents: A Comparative Study / E. Konstantinova, Y. Shalaumova, T. Maslakova [et al.] // *International J. of Medical Research & Health Sciences*. – 2018. – Vol. 7, No. 3. – P. 1–7.
17. Mosteller, F. Data analysis and regression. A second course in statistics / F. Mosteller, J. W. Tukey. – Addison-Westly Publishing Company, 1978. – 317 p. – ISBN 978-0201048544.
18. Siegel, A. F. Practical Business Statistics, 7<sup>th</sup> ed. / A. F. Siegel. – Academic Press, 2016. – 642 p. – ISBN 978-0128042502.
19. Гмурман, В. Е. Теория вероятностей и математическая статистика : учеб. 12-е изд. / В. Е. Гмурман. – М. : Издательство Юрайт, 2020. – 479 с. – (Профессиональное образование). – ISBN 978-5-534-00859-3. – EDN DIXQIX.
20. Sweet, S. A. Data Analysis with SPSS. A First Course in Applied Statistics, 4<sup>th</sup> ed. / S. A. Sweet, K. Grace-Martin. – Pearson, 2010. – 288 p. – ISBN-13: 978-0205019670.
21. Боровиков, В. STATISTICA: искусство анализа данных на компьютере. Для профессионалов / В. Боровиков. – СПб.: Питер, 2001. – 656 с. – ISBN 5-318-00302-8.
22. Norušis, M. J. SPSS 16.0 Guide to Data Analysis / M. J. Norušis. – Prentice Hall, 2008. – 653 p.

23. *Nannan, N.* Estimating completeness of birth registration in South Africa, 1996–2011 / N. Nannan, R. Dorrington, D. Bradshaw // *Bull World Health Organ.* – 2019. – No. 97. – P. 468–476. – DOI 10.2471/BLT.18.222620.
24. *Hoekstra, R.* Are assumptions of well-known statistical techniques checked, and why (not)? / R. Hoekstra, H. A. L. Kiers, A. Johnson // *Front. Psychology.* – 2012. – Vol. 3, No. 137. – DOI 10.3389/fpsyg.2012.00137.
25. *Dye, C.* Tuberculosis decline in populations affected by HIV: a retrospective study of 12 countries in the WHO African Region / C. Dye, B. G. Williams // *Bull World Health Organ.* – 2019. – No. 97. – P. 405–414. – DOI: 10.2471/BLT.18.228577.
26. *Byass, P.* Correlation between noncommunicable disease mortality in people aged 30–69 years and those aged 70–89 years / P. Byass // *Bull World Health Organ.* – 2019. – No. 97. – P. 589–596. – DOI 10.2471/BLT.18.227132.
27. *Урбах, В. Ю.* Биометрические методы. Статистическая обработка опытных данных в биологии, сельском хозяйстве и медицине / В. Ю. Урбах. – М. : Наука, 1964. – 416 с.
28. Свидетельство о государственной регистрации программы для ЭВМ № 2019612871 Российская Федерация. Программа для расчета скользящего среднего и проверки условий применимости моделей регрессии : № 2019611507 : заявл. 18.02.2019 : опубл. 04.03.2019 / Ю. В. Шалаумова, А. Н. Вараксин, В. Г. Панов; заявитель Федеральное государственное бюджетное учреждение науки Институт промышленной экологии Уральского отделения Российской академии наук (ИПЭ УрО РАН). – EDN MVJCBS.
29. Применение методов скользящего среднего для построения регрессионных моделей в медико-экологических исследованиях / А. Н. Вараксин, Ю. В. Шалаумова, Т. А. Маслакова [и др.] // *Экологические системы и приборы.* – 2020. – № 6. – С. 12–21. – DOI 10.25791/esip.06.2020.1159. – EDN XTBFVAV.
30. *Hosmer, D.* Applied Logistic Regression, 2<sup>nd</sup> ed. / D. Hosmer, S. Lemeshow. – New York: Wiley, 2000. – 397 p. – ISBN 0-471-35632-8.
31. *Shoukri, M. M.* Statistical methods for health sciences, 2<sup>nd</sup> ed. / M. M. Shoukri, C. A. Pause. – CRC Press, 1998. – 384 p. – ISBN 978-0849310959.
32. *Cox, D. R.* The analysis of binary data, 2<sup>nd</sup> ed. / D. R. Cox, E. J. Snell. – London: Chapman and Hall, 1989. – 240 p. – ISBN 9780412306204.
33. *Nagelkerke, N. J. D.* A note on a general definition of the coefficient of determination / N. J. D. Nagelkerke // *Biometrika.* – 1991. – Vol. 78, No. 3. – P. 691–692. – DOI 10.1093/biomet/78.3.691.
34. *Peng, J.* An Introduction to Logistic Regression Analysis and Reporting / J. Peng, K. L. Lee, G. M. Ingersoll // *J. of Educational Research.* – 2002. – Vol. 96, No. 1. – P. 3–14. – DOI 10.1080/00220670209598786. – EDN EGUCMX.
35. Эрозия слизистой оболочки влагалища у женщин с хирургической коррекцией пролапса гениталий в постменопаузе / А. А. Михельсон, М. В. Лазукина, А. Н. Вараксин [и др.] // *Лечение и профилактика.* – 2020. – Т. 10, № 4. – С. 55–64. – EDN ZCTUDM.
36. Fluid balance, intradialytic hypotension, and outcomes in critically ill patients undergoing renal replacement therapy: a cohort study / J. A. Silversides, R. Pinto, R. Kuint [et al.] // *Critical Care.* – 2014. – Vol. 18, No. 6. – DOI 10.1186/s13054-014-0624-8.
37. Hearing and vision screening for preschool children using mobile technology, South Africa / S. Eksteen, S. Launer, H. Kuper [et al.] // *Bull World Health Organ.* – 2019. – No. 97. – P. 672–680. – DOI 10.2471/BLT.18.227876.
38. *Тьюки, Дж.* Анализ результатов наблюдений / Дж. Тьюки. – М. : Мир, 1981. – 695 с.

## **Сведения об авторах:**

**Вараксин Анатолий Николаевич**, д. ф.-м. н., профессор, главный научный сотрудник Института промышленной экологии УрО РАН, г. Екатеринбург, ул. С. Ковалевской, 20, Россия. Эл. почта: varaksin@esko.uran.ru.

**Шалаумова Юлия Валерьевна**, канд. т. н., научный сотрудник.

**Маслакова Татьяна Анатольевна**, канд. ф.-м. н., научный сотрудник.

# LINEAR AND NONLINEAR REGRESSION IN BIOLOGY AND MEDICINE: A MOVING AVERAGE APPROACH

A. N. Varaksin, Yu. V. Shalaumova, T. A. Maslakova

*Institute of Industrial Ecology, Ural Branch of Russian Academy of Sciences,  
Ekaterinburg, Russia*

*Testing of assumptions of linear and non-linear (logistic) regressions are discussed in this method paper. We propose a moving average approach as one of the methods for checking the linearity of the relationship between predictor  $X$  and response  $Y$  in linear regression and the linearity of the relationship between predictor  $X$  and the logit of response  $Y$  in logistic regression. Specific examples show the advantages of a moving average over the common data stratification approach. The importance of graphical representation of moving average results in linear and (especially) logistic regression is emphasized.*

**Key words:** moving average; linearity test; linear regression; logistic regression.

## References

1. Osborne, J. W. Multiple Regression Assumptions / J. W. Osborne, E. Waters. – ERIC Digest, 2002.
2. Park, C. C. A linearity test statistic in a simple linear regression / C. C. Park, K. E. Lee // J. of the Korean Data and Information Science Society. – 2014. – Vol. 25, No. 2. – P. 305–315. – DOI 10.7465/jkdi.2014.25.2.305.
3. Nimon, K. F. Statistical assumptions of substantive analyses across the general linear model: a mini-review / K. F. Nimon // Front. Psychology. – 2012. – Vol. 3, No. 322. – DOI 10.3389/fpsyg.2012.00322.
4. Dhakal, C. P. Regression invented as statistics / C. P. Dhakal // International J. of Interdisciplinary Research and Innovations. – 2018. – Vol. 6, No. 2. – P. 1–5.
5. Stanton, J. M. Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors / J. M. Stanton // J. of Statistics Education. – 2001. – Vol. 9, No. 3. – DOI 10.1080/10691898.2001.11910537.
6. Aldrich, J. Fisher and regression / J. Aldrich // Statistical Science. – 2005. – Vol. 20, No. 4. – P. 401–417.
7. Draper, N. R. Applied Regression Analysis, 3<sup>rd</sup> ed. / N. R. Draper, H. Smith. – Wiley-Interscience Publication, 1998. – 736 p. – ISBN 978-0-471-17082-2.
8. Afifi, A. A. Statistical Analysis: A Computer Oriented Approach / A. A. Afifi, S. P. Azen. – New York : Academic Press, 1972. – 366 p. – ISBN: 9780120444502.
9. Afifi, A. A. Computer-aided multivariate analysis, 4<sup>th</sup> ed. / A. A. Afifi, S. May, V. Clark. – Chapman&Hall/CRC, 2003. – 512 p. – ISBN 978-1584883081.



10. An Analysis of the Links between Smoking and BMI in Adolescents: A Moving Average Approach to Establishing the Statistical Relationship between Quantitative and Dichotomous Variables / A. N. Varaksin, E. D. Konstantinova, T. A. Maslakova [et al.] // *Children (Basel)*. – 2022. – Vol. 9, No. 2. – DOI 10.3390/children9020220. – EDN GKBQJD.
11. *Vladimirov, V. G.* Radioprotective formulations. Optimization of composition and mechanism of action / V. G. Vladimirov, G. A. Poddubsky, G. I. Razorenov. – Leningrad : Voenizdat, 1988. – 144 p. (In Russian).
12. *Nalimov, V. V.* Theory of experiment / V. V. Nalimov. – Moscow : Nauka, 1971. – 208 p. (In Russian).
13. Are in vivo and in vitro assessments of comparative and combined toxicity of the same metallic nanoparticles compatible, or contradictory, or both? A juxtaposition of data obtained in respective experiments with NiO and Mn<sub>3</sub>O<sub>4</sub> nanoparticles / I. A. Minigalieva, T. V. Bushueva, E. Fröhlich [et al.] // *Food and Chemical Toxicology*. – 2017. – Vol. 109, No. 1. – P. 393–404. – DOI 10.1016/j.fct.2017.09.032. – EDN ZGWVNH.
14. *Fletcher, R. H.* Clinical epidemiology: The essentials, 5<sup>th</sup> ed. / R. H. Fletcher, S. W. Fletcher. – Lippincott Williams & Wilkins, 2012. – 272 p. – ISBN 978-1451144475.
15. *Rasina, L. N.* Statistical analysis of ecological and physiological data in the interpretation of technogenic impacts / L. N. Rasina, N. A. Orekhova, A. N. Varaksin // *Ecological systems and devices*. – 2011, No. 12. – P. 50–53. – EDN SHXDFX. (In Russian).
16. Effects of Environmental Radioactive Pollution on the Cardiovascular Systems of Ural Region Residents: A Comparative Study / E. Konstantinova, Y. Shalaumova, T. Maslakova [et al.] // *International J. of Medical Research & Health Sciences*. – 2018. – Vol. 7, No. 3. – P. 1–7.
17. *Mosteller, F.* Data analysis and regression. A second course in statistics / F. Mosteller, J. W. Tukey. – Addison-Westly Publishing Company, 1978. – 317 p. – ISBN 978-020-1048544.
18. *Siegel, A. F.* Practical Business Statistics, 7<sup>th</sup> ed. / A. F. Siegel. – Academic Press, 2016. – 642 p. – ISBN 978-0128042502.
19. *Gmurman, V. E.* Probability theory and mathematical statistics: textbook, 12<sup>th</sup> ed. / V. E. Gmurman. – Moscow : Yurayt Publishing House, 2020. – 479 p. – (Professional education). – ISBN 978-5-534-00859-3. – EDN DIXQIX. (In Russian).
20. *Sweet, S. A.* Data Analysis with SPSS. A First Course in Applied Statistics, 4<sup>th</sup> ed. / S. A. Sweet, K. Grace-Martin. – Pearson, 2010. – 288 p. – ISBN-13: 978-0205019670.
21. *Borovikov, V.* STATISTICA: the art of computer data analysis. For professionals / V. Borovikov. – St. Petersburg: Piter, 2001. – 656 p. – ISBN 5-318-00302-8.
22. *Norušis, M. J.* SPSS 16.0 Guide to Data Analysis / M. J. Norušis. – Prentice Hall, 2008. – 653 p.
23. *Nannan, N.* Estimating completeness of birth registration in South Africa, 1996–2011 / N. Nannan, R. Dorrington, D. Bradshaw // *Bull World Health Organ*. – 2019. – No. 97. – P. 468–476. – DOI 10.2471/BLT.18.222620.
24. *Hoekstra, R.* Are assumptions of well-known statistical techniques checked, and why (not)? / R. Hoekstra, H. A. L. Kiers, A. Johnson // *Front. Psychology*. – 2012. – Vol. 3, No. 137. – DOI 10.3389/fpsyg.2012.00137.
25. *Dye, C.* Tuberculosis decline in populations affected by HIV: a retrospective study of 12 countries in the WHO African Region / C. Dye, B. G. Williams // *Bull World Health Organ*. – 2019. – No. 97. – P. 405–414. – DOI: 10.2471/BLT.18.228577.

26. *Byass, P.* Correlation between noncommunicable disease mortality in people aged 30–69 years and those aged 70–89 years / P. Byass // *Bull World Health Organ.* – 2019. – No. 97. – P. 589–596. – DOI 10.2471/BLT.18.227132.
27. *Urbah, V. Yu.* Biometric methods. Statistical processing of experimental data in biology, agriculture and medicine. – Moscow : Nauka, 1964. – 416 p. (In Russian).
28. Certificate of state registration of the computer program No. 2019612871 Russian Federation. Program for calculating the moving average and checking the conditions for the applicability of regression models: No. 2019611507: Appl. 02/18/2019 : publ. 03/04/2019 / Yu. V. Shalaumova, A. N. Varaksin, V. G. Panov; applicant Federal State Budgetary Institution of Science Institute of Industrial Ecology of the Ural Branch of the Russian Academy of Sciences (IIE Ural Branch of the Russian Academy of Sciences). – EDN MVJCBS. (In Russian).
29. Application of moving average methods for constructing regression models in medical and environmental studies / A. N. Varaksin, Yu. V. Shalaumova, T. A. Maslakova [et al.] // *Ecological systems and devices.* – 2020, No. 6. – P. 12–21. – DOI 10.25791/esip.06.2020.1159. – EDN XTBFVAV. (In Russian).
30. *Hosmer, D.* Applied Logistic Regression, 2<sup>nd</sup> ed. / D. Hosmer, S. Lemeshow. – New York: Wiley, 2000. – 397 p. – ISBN 0-471-35632-8.
31. *Shoukri, M. M.* Statistical methods for health sciences, 2<sup>nd</sup> ed. / M. M. Shoukri, C. A. Pause. – CRC Press, 1998. – 384 p. – ISBN 978-0849310959.
32. *Cox, D. R.* The analysis of binary data, 2<sup>nd</sup> ed. / D. R. Cox, E. J. Snell. – London: Chapman and Hall, 1989. – 240 p. – ISBN 9780412306204.
33. *Nagelkerke, N. J. D.* A note on a general definition of the coefficient of determination / N. J. D. Nagelkerke // *Biometrika.* – 1991. – Vol. 78, No. 3. – P. 691–692. – DOI 10.1093/biomet/78.3.691.
34. *Peng, J.* An Introduction to Logistic Regression Analysis and Reporting / J. Peng, K. L. Lee, G. M. Ingersoll // *J. of Educational Research.* – 2002. – Vol. 96, No. 1. – P. 3–14. – DOI 10.1080/00220670209598786.– EDN EGUCMX.
35. Erosion of the vaginal mucosa in postmenopausal women with surgical correction of genital prolapse / A. A. Mikhelson, M. V. Lazukina, A. N. Varaksin [et al.] // *Treatment and prevention.* – 2020. – Vol. 10, No. 4. – P. 55–64. – EDN ZCTUDM. (In Russian).
36. Fluid balance, intradialytic hypotension, and outcomes in critically ill patients undergoing renal replacement therapy: a cohort study / J. A. Silversides, R. Pinto, R. Kuint [et al.] // *Critical Care.* – 2014. – Vol. 18, No. 6. – DOI 10.1186/s13054-014-0624-8.
37. Hearing and vision screening for preschool children using mobile technology, South Africa / S. Eksteen, S. Launer, H. Kuper [et al.] // *Bull World Health Organ.* – 2019. – No. 97. – P. 672–680. – DOI 10.2471/BLT.18.227876.
38. *Tukey, J.* Analysis of observational results / J. Tukey. – Moscow : Mir, 1981. – 695 p.

# РАЗРАБОТКА МЕТОДА УПРАВЛЕНИЯ АКТИВНЫМ ЭКЗОСКЕЛЕТОМ ДЛЯ РЕАБИЛИТАЦИИ МОТОРНОЙ ДЕЯТЕЛЬНОСТИ ВЕРХНИХ КОНЕЧНОСТЕЙ ЧЕЛОВЕКА

В. М. Антипов<sup>1,2</sup>, А. А. Бадарин<sup>1,3</sup>, А. В. Андреев<sup>1</sup>, В. В. Грубов<sup>1,3</sup>, Е. Н. Пицик<sup>1,3,4</sup>, С. А. Лобов<sup>1</sup>, В. А. Максименко<sup>1,4,5</sup>, В. Б. Казанцев<sup>1,3</sup>, А. Е. Храмов<sup>1,3,4</sup>

<sup>1</sup> Федеральное государственное автономное образовательное учреждение высшего образования «Балтийский федеральный университет им. Иммануила Канта», Балтийский центр нейротехнологий и искусственного интеллекта, г. Калининград, Россия

<sup>2</sup> Радиофизический факультет, Нижегородский государственный университет им. Лобачевского, г. Н. Новгород, Россия

<sup>3</sup> НИИ Нейронаук, Самарский государственный медицинский университет, г. Самара, Россия

<sup>4</sup> Уральский Федеральный университет им. первого Президента России Б. Н. Ельцина, г. Екатеринбург, Россия

<sup>5</sup> Автономная некоммерческая организация высшего образования «Университет Иннополис», г. Казань, Россия

*В настоящей работе представлено описание результатов разработки методов управления активным экзоскелетом для реабилитации моторной деятельности верхних конечностей человека. Описана экспериментальная установка, программная реализация системы управления в реальном времени, а также разработанная программная среда для стимуляции испытуемого в процессе проведения нейрофизиологического эксперимента. Объектом исследования описанной системы являются процессы активации участков мозга человека с точки зрения анализа регистрируемых биоэлектрических электроэнцефалографических сигналов во время двигательной активности, пригодные для использования в задачах реабилитации двигательной активности. Цель работы заключается в разработке программно-аппаратной части эксперимента по исследованию алгоритмов реального времени для управления экзоскелетом верхних конечностей для задач реабилитации с использованием интерфейсов «мозг-компьютер».*

**Ключевые слова:** экзоскелет; приводы; интерфейс «мозг-компьютер»; тренировка моторной деятельности.

## 1. Введение

Традиционно одна из основных областей применения интерфейсов «мозг-компьютер» (ИМК) – управление внешними устройствами, например, антропоморфными манипуляторами, роботами и экзоскелетами, с помощью нейронной активности, генерируемой воображаемыми движениями [1]. В связи с этим применение

ИМК направлено на улучшение качества жизни людей с нарушением двигательных функций вследствие травм, нейродегенеративных заболеваний или потери конечностей. В частности, большое количество работ посвящено разработке интерфейсов «мозг-компьютер» для реабилитации больных после инсульта [2] и травмы спинного мозга [3], а также для управления биопротезом после ампутации конечности [4] и паралича [5].

## 2. Экспериментальная установка

Для проведения эксперимента по исследованию алгоритмов реального времени для управления экзоскелетом был выбран двухсторонний нейротренажер верхних конечностей 2N (производитель ООО «Нейроассистивные технологии», Россия). Основные компоненты в используемом нейротренажере – две роботизированные руки, оснащенные шестью активными и шестью пассивными степенями свободы. Основные компоненты и внешний вид экспериментальной установки изображены на рис. 1.

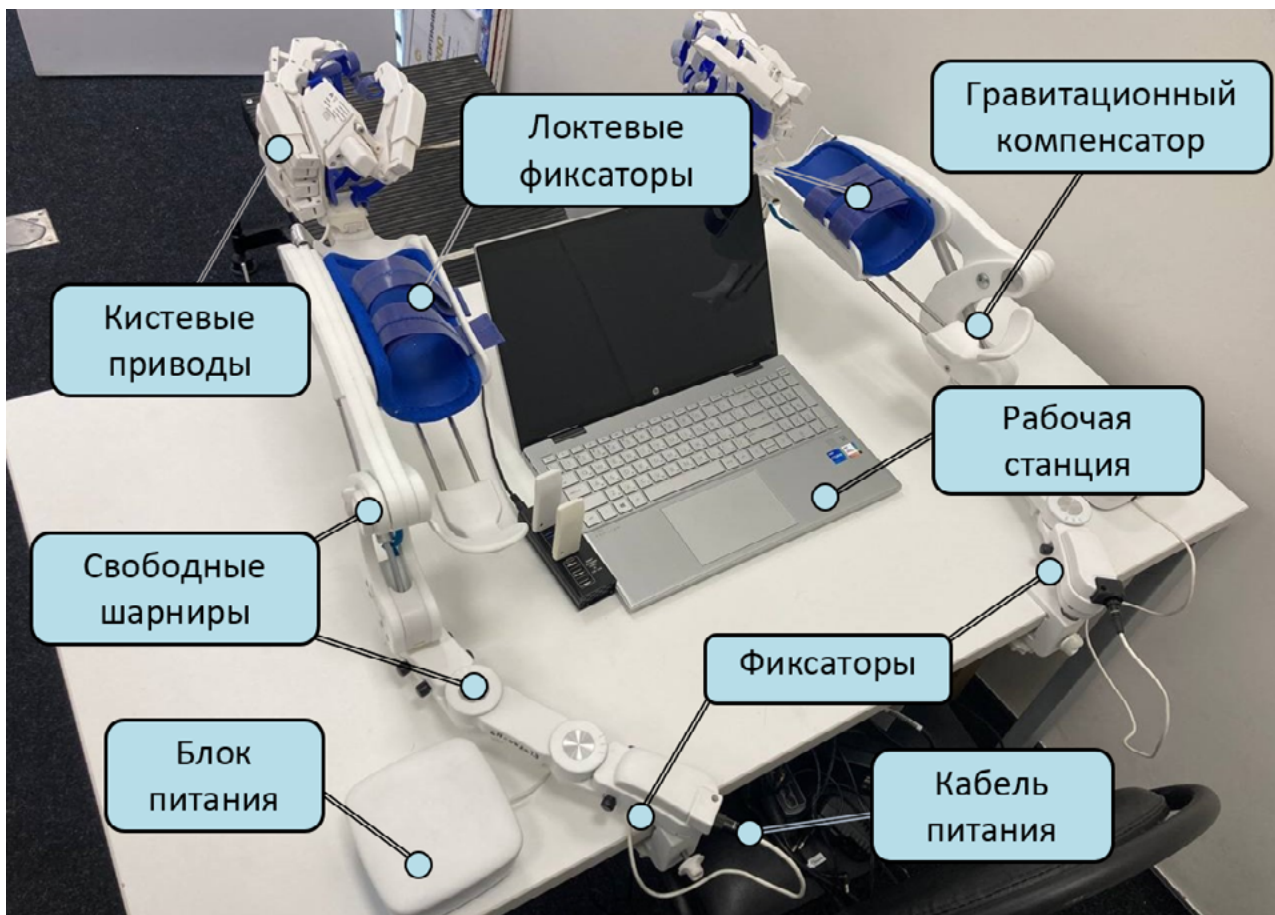


Рис. 1. Экспериментальная установка

В каждом активном шарнире установлены линейные сервоприводы, которые позволяют сгибать и разгибать отдельные пальцы кисти руки человека. Установка линейных редукторов в сервоприводах обусловлена обеспечением высокого момента на выходном валу исполнительного механизма. Это позволяет преодолевать мышечное сопротивление в случае спастики или если испытуемый полностью или частично утратил двигательные функции. Каждый палец надежно фиксируется с помощью гибких самоклеящихся липучек (рис. 2).

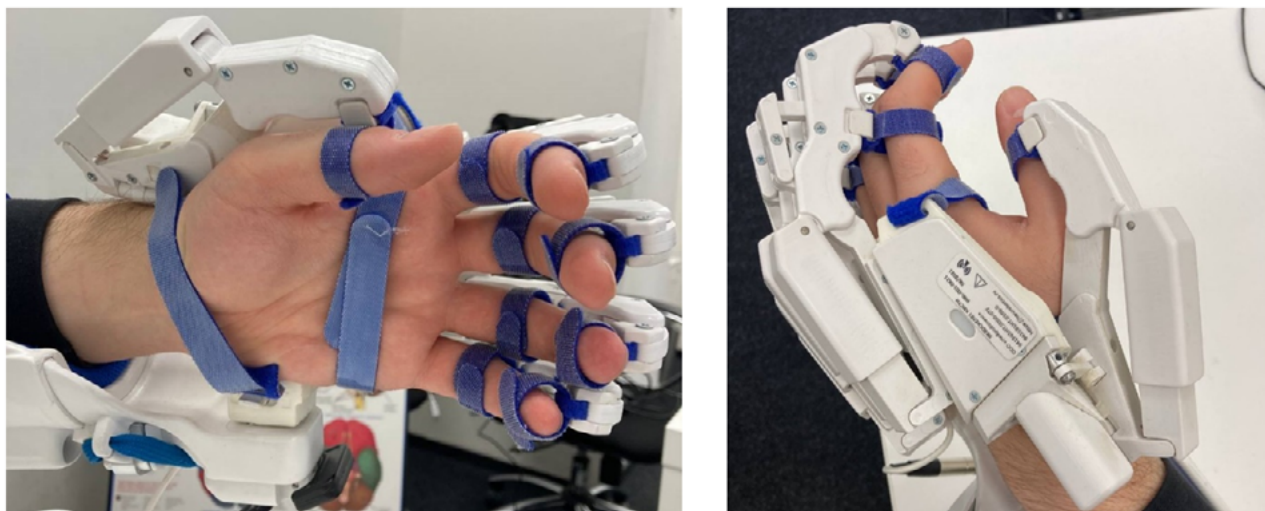


Рис. 2. Фиксаторы кисти человека

Фиксация позволяет гарантировать правильное положение пальцев рук, а также правильность выполнения упражнения. Также нейротренажер оснащен дополнительным приводом для позиционирования большого пальца во фронтальной плоскости; это позволяет улучшить эргономичность крепления и адаптировать устройство к индивидуальным особенностям испытуемого. Для обеспечения более надежной фиксации всей руки используются локтевые фиксаторы (рис. 1). Они представляют собой более широкие липкие ленты, закрепленные на подлокотнике. Устройство имеет выдвижной упор, который позволяет точнее установить точку опоры локтевого сустава с учетом длины рук испытуемого.

Пассивные (свободные) шарниры предназначены для обеспечения полного набора естественных движений человека во время прохождения эксперимента. В локтевом шарнире дополнительно установлен «гравитационный компенсатор», который позволяет компенсировать вес всей руки для разгрузки плечевых мышц человека. Это дает возможность испытуемому более комфортно переносить длительные исследования. Каждый пассивный шарнир обладает возможностью блокировки в определенном положении с помощью фиксаторов, изображенных на рис. 3. Эта возможность позволяет учесть специфику каждого отдельного эксперимента, например, если отдельно задействована только кисть руки.



Рис. 3. Расположение основных фиксаторов нейротренажера для блокировки пассивных степеней свободы



Одним из естественных движений человека является вращение кисти руки в сагиттальной плоскости. Для обеспечения данного вращения в нейротренажере установлен сагиттальный шарнир, представленный на рис. 3. Он также имеет фиксатор для блокировки необходимого положения и может быть использован для более удобного позиционирования кисти руки.

Обе руки можно задействовать как совместно, так и отдельно, предварительно установив необходимое программное обеспечение на рабочую станцию. В качестве рабочей станции может выступать обычный персональный компьютер с установленной операционной системой Windows. Одним из основных требований к рабочей станции является наличие необходимого количества USB Type A портов для подключения всех внешних модулей для управления устройством. В расширенной версии устройства могут потребоваться дополнительные порты ввода/вывода.

Для управления каждой отдельно взятой рукой нейротренажера используется кастомный модуль беспроводной связи. Данный модуль позволяет в реальном времени и на большом расстоянии отправлять и получать управляющие команды с устройства. Для работы с данным модулем необходимо установить дополнительные драйверы.

Для повышения мобильности нейротренажер оснащен внешним аккумуляторным блоком. Данный блок позволяет устройству работать в непрерывном режиме более 4 ч.

Аккумуляторные блоки полностью совместимы и подходят как для правой, так и для левой руки, что позволяет проводить длинные эксперименты, не прерываясь на подзарядку.

### 3. Система автоматического управления в реальном времени

Для управления роботизированными руками нейротренажера в реальном времени был разработан программный модуль, позволяющий контролировать положение каждого привода системы. Общая функциональная схема работы экспериментальной установки, а также методы взаимодействия между различными блоками и сопрограммами системы изображены на рис. 4.

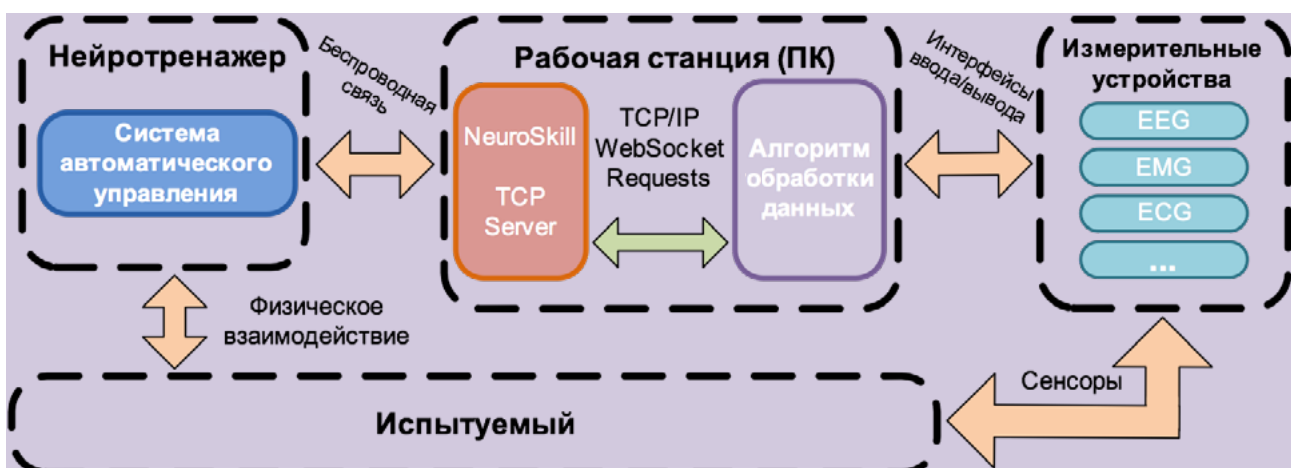


Рис. 4. Функциональная схема работы экспериментальной установки, взаимодействие и способы передачи данных между различными блоками системы

Управление сервоприводами на низком уровне осуществляется с помощью САУ, установленной на контроллерах каждой роботизированной руки. Контроллеры,

в свою очередь, подключаются посредством беспроводной связи к рабочей станции с помощью специальных модулей. Основной функционал по работе с сетевыми интерфейсами обеспечивает программное обеспечение NeuroSkill. Данная программа развертывает поверх TCP/IP стека веб-сервер в локальной сети. Для доступа к серверу предоставляется API, основанное на WebSocket (http) запросах. Для работы с данным API был разработан программный модуль – ВрAPPER на языке Python.

Данный модуль представляет собой класс, предоставляющий программный интерфейс для работы с каждым активным приводом нейротренажера. Набор методов позволяет гибко управлять скоростью, силой сжатия/разжатия и положением каждой конечности кисти руки. Для детектирования правильности выполнения команд программа позволяет в режиме реального времени получать текущее состояние различных параметров системы. Для этого отправляется специальный GET-запрос на сервер, после чего в ответ приходит JSON-структура, которая описывает запрашиваемые состояния. Парсинг данной структуры и чтение каждого ее поля позволяют получить нужный параметр и скорректировать выполнение основной программы.

Для обеспечения работы в реальном времени ВрAPPER запускается в отдельном потоке, что позволяет поддерживать высокую скорость работы и низкую латентность системы. Это обусловлено высокой нагрузкой на основной модуль программы, занимающийся обработкой и анализом данных с ЭЭГ. Для измерения ЭЭГ-сигналов используется электроэнцефалограф ActiChamp, который позволяет считывать 64 канала сенсорной информации на частоте выше 1000 Гц. Сенсоры закрепляются на голове пациента с помощью индивидуально подобранной шапки acti-Cap. Кроме того, на каждой руке можно установить по 3 сенсора ЭМГ, запись с которых производится с помощью модуля «Энцефалан-ЭЭГР-19/27» с частотой дискретизации 250 Гц.

Исходя из этого, разработанную функциональную схему экспериментальной установки можно описать следующим образом: на испытуемого закрепляются различные сенсоры, записывающие биоэлектрические сигналы; сенсоры подключаются к измерительному устройству (усилителю), которое проводит первичную фильтрацию и буферизацию измеренной информации. Затем измерительные устройства подключаются к рабочей станции (например, персональный компьютер, одноплатный компьютер и т. п.) и передают измеренные данные в рабочую программу BrainVision (или аналогичную), которая ретранслирует данные в модуль Алгоритма обработки данных. Алгоритм реализует методы выделения паттернов, ассоциированных с воображаемыми движениями верхними конечностями.

Далее посредством WebSocket запросов управляющие воздействия передаются на программное обеспечение NeuroSkill, которое преобразует и ретранслирует необходимые данные через беспроводной модуль связи в САУ нейротренажера. После этого устройство начинает выполнять упражнение и активировать руки испытуемого путем сгибательно-разгибательных движений, замыкая тем самым контур биологической обратной связи, что позволяет повысить эффективность процесса реабилитации моторной деятельности верхних конечностей человека с использованием активного экзоскелета.

#### **4. Стимуляционная среда**

Процесс проведения сложных нейрофизиологических экспериментов требует наличия специальных программных инструментов, позволяющих выполнять различные нетривиальные задачи и реализовывать таким образом сценарий экспери-



мента, что становится особенно важным при решении задач нейрореабилитации. Следует отметить, что каждый отдельно взятый эксперимент требует точного соблюдения временных интервалов, своевременного ведения лога данных, а также иметь возможность синхронизации с различными внешними устройствами. Для решения поставленных задач нами было разработано программное обеспечение Neuro-Physiological Experiments Studio (далее – NPESstudio). Данная программа позволяет в режиме реального времени предъявлять испытуемому различные стимулы, контролировать течение эксперимента, синхронизироваться с внешними измерительными устройствами, обрабатывать внешние триггеры с портов ввода/вывода, а также обеспечивает возможность получения обратной связи через сенсорные дисплеи и устройства ввода информации. Внешний вид основного окна описанной программы представлен на рис. 5.

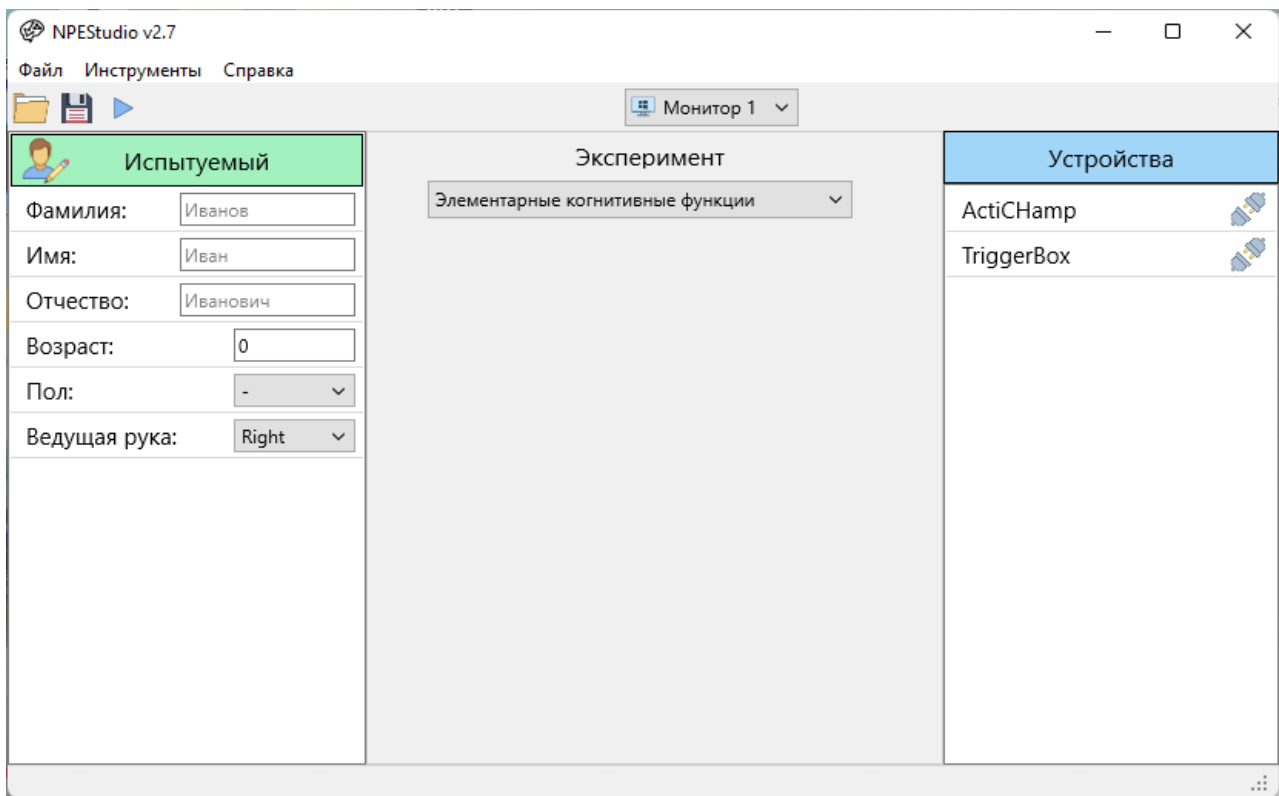


Рис. 5. Главное окно программы NPESstudio, где: «Испытуемый» – поля ввода основной информации о человеке, проходящем эксперимент; «Эксперимент» – элемент управления, позволяющий выбирать экспериментальную программу; «Устройства» – панель статуса подключения внешних устройств

Программа разработана на кроссплатформенном фреймворке FireMonkey в программной среде Embarcadero RAD Studio на языке программирования C++. Данный фреймворк позволяет запускать разработанную программу на таких операционных системах, как Windows, Linux, Android, MacOS, IOS. Ключевой особенностью FireMonkey является использование единой базы исходного кода, который легко компилируется на различные платформы.

Основное ядро разработанного приложения занимает модуль (класс) когнитивных задач, или Cognitive Test Stack. Данный модуль представляет собой масштабируемый контейнер когнитивных тестов, где каждый тест имеет обособленный модуль настроек с возможностью подключения к внешним измеряемым устройствам через управляющие потоки ввода/вывода. Для сохранения и стриминга данных

в программе реализована поддержка LSL протокола. Lab Streaming Layer (далее – LSL) – это система для унифицированного сбора временных рядов измерений в исследовательских экспериментах, синхронизации и доступа в реальном времени, а также для централизованного сбора, просмотра и записи данных на диск. Для удобного хранения и обработки данных в программе реализована поддержка формата mat.

Основные модули программы встраиваются в иерархичную структуру наследования классов, что позволяет легко наследовать и масштабировать отдельные модули для специфичных задач. Каждая задача (эксперимент), унаследованная от базового класса BaseTask, может легко встраиваться в общий стек задач. Это позволяет без особых усилий проектировать новые нейрофизиологические эксперименты. Также данный класс отвечает за отрисовку и визуализацию стимулов. Подобным образом реализованы все основные модули программы, например, триалы (совокупности повторяющихся стимулов).

## 5. Результаты

В ходе выполнения работ были разработан метод управления активным экзоскелетом для реабилитации моторной деятельности верхних конечностей человека, а также кроссплатформенное программное обеспечение, позволяющее быстро реализовывать различные мультимодальные алгоритмы стимуляции испытуемого в процессе проведения различных нейрофизиологических экспериментов.

## 6. Благодарности

Исследование выполнено при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках Программы развития Балтийского федерального университета имени И. Канта в соответствии с программой стратегического академического лидерства «Приоритет-2030».

## 7. Список литературы

1. He, B. Noninvasive brain-computer interfaces based on sensorimotor rhythms / B. He, B. Baxter, B. J. Edelman [et al.] // Proceedings of the IEEE. – 2015. – Vol. 103, No. 6. – P. 907–925. – DOI 10.1109/JPROC. 2015.2407272.
2. Ang, K. K. Brain-computer interface for neurorehabilitation of upper limb after stroke / K. K. Ang, C. Guan // Proceedings of the IEEE. – 2015. – Vol. 103, No. 6. – P. 944–953. – DOI 10.1109/JPROC. 2015.2415800.
3. Müller-Putz, G. R. Moregrasp: Restoration of upper limb function in individuals with high spinal cord injury by multimodal neuroprostheses for interaction in daily activities / G. R. Müller-Putz, P. Ofner, A. Schwarz [et al.] // 7<sup>th</sup> Graz Brain-Computer Interface Conference. – 2017. – Vol. 7. – P. 338–343. – DOI 10.3217/978-3-85125-533-1-62.
4. Murphy, D. P. Electroencephalogram-based brain-computer interface and lower-limb prosthesis control: A case study / D. P. Murphy, O. Bai, A. S. Gorgey [et al.] // Front. Neurol. – 2017. – Vol. 8. – P. 696. – DOI 10.3389/fneur.2017.00696.
5. Bowsher, K. Brain-computer interface devices for patients with paralysis and amputation: a meeting report / K. Bowsher, E. F. Civillico, J. Coburn [et al.] // J. Neural Eng. – 2016. – Vol. 13, No. 2. – P. 23001. – DOI 10.1088/1741-2560/13/2/023001.

## Сведения об авторах:

**Антипов Владимир Михайлович**, аспирант, радиофизический факультет, Нижегородский государственный университет им. Лобачевского, г. Н. Новгород, Россия. Эл. почта: vantipovm@gmail.com.

**Бадарин Артем Александрович**, к. ф.-м. н., старший научный сотрудник Балтийского центра нейротехнологий и искусственного интеллекта, Балтийский федеральный университет им. И. Канта, ул. А. Невского, 14, г. Калининград, 236041, Россия, старший научный сотрудник НИИ Нейронаук, Самарский государственный медицинский университет, 443079, г. Самара, ул. Гагарина, 18, Россия. Эл. почта: badarin.a.a@mail.ru

**Андреев Андрей Викторович**, к. ф.-м. н., старший научный сотрудник Балтийского центра нейротехнологий и искусственного интеллекта, Балтийский федеральный университет им. И. Канта, ул. А. Невского, 14, г. Калининград, 236041, Россия. Эл. почта: andreevandreii1993@gmail.com

**Грубов Вадим Валерьевич**, к. ф.-м. н., старший научный сотрудник Балтийского центра нейротехнологий и искусственного интеллекта, Балтийский федеральный университет им. И. Канта, ул. А. Невского, 14, г. Калининград, 236041, Россия, старший научный сотрудник НИИ Нейронаук, Самарский государственный медицинский университет, 443079, г. Самара, ул. Гагарина, 18, Россия. Эл. почта: vvgrubov@gmail.com

**Пицик Елена Николаевна**, младший научный сотрудник Балтийского центра нейротехнологий и искусственного интеллекта, Балтийский федеральный университет им. И. Канта, ул. А. Невского, 14, г. Калининград, 236041, Россия, младший научный сотрудник НИИ Нейронаук, Самарский государственный медицинский университет, 443079, г. Самара, ул. Гагарина, 18, Россия. Эл. почта: pitsikelena@gmail.com

**Лобой Сергей Анатольевич**, д. ф.-м. н., ведущий научный сотрудник Балтийского центра нейротехнологий и искусственного интеллекта, Балтийский федеральный университет им. И. Канта, ул. А. Невского, 14, г. Калининград, 236041, Россия. Эл. почта: lobov@neuro.nnov.ru

**Максименко Владимир Александрович**, д. ф.-м. н., ведущий научный сотрудник Балтийского центра нейротехнологий и искусственного интеллекта, Балтийский федеральный университет им. И. Канта, ул. А. Невского, 14, г. Калининград, 236041, Россия. Эл. почта: maximenkovl@gmail.com

**Казанцев Виктор Борисович**, д. ф.-м. н., ведущий научный сотрудник Балтийского центра нейротехнологий и искусственного интеллекта, Балтийский федеральный университет им. И. Канта, ул. А. Невского, 14, г. Калининград, 236041, Россия, руководитель лаборатории НИИ Нейронаук, Самарский государственный медицинский университет, 443079, г. Самара, ул. Гагарина, 18, Россия. Эл. почта: kazantsev@neuro.nnov.ru

**Храмов Александр Евгеньевич**, д. ф.-м. н., профессор, главный научный сотрудник Балтийского центра нейротехнологий и искусственного интеллекта, Балтийский федеральный университет им. И. Канта, ул. А. Невского, 14, г. Калининград, 236041, Россия, руководитель лаборатории НИИ Нейронаук, Самарский государственный медицинский университет, 443079, г. Самара, ул. Гагарина, 18, Россия. Эл. почта: a.hramov@innopolis.ru.

## DEVELOPMENT OF A METHOD FOR CONTROLLING AN ACTIVE EXOSKELETON FOR THE REHABILITATION OF MOTOR ACTIVITY OF THE HUMAN UPPER LIMBS

V. M. Antipov, A. A. Badarin, A. V. Andreev, V. V. Grubov, E. N. Pitsik, S. A. Lobov,  
V. A. Maksimenko, V. B. Kazantsev, A. E. Hramov

*Neuroscience and Cognitive Technology Laboratory, Center for Technologies in Robotics  
and Mechatronics Components, Innopolis University, Innopolis, Russia*

*We present the results of the development of methods for controlling an active exoskeleton for the rehabilitation of motor activity of the human upper limbs. We describe an experimental setup, a software implementation of a real-time control system, and a developed software environment for stimulating a subject during a neurophysiological experiment. The object of study of the described system is the processes of activation of human brain regions from the point of view of the analysis of recorded bioelectrical electroencephalographic signals during motor activity, suitable for use in the tasks of rehabilitation of motor activity. The purpose of the work is to develop the software and hardware part of the experiment to study real-time algorithms for controlling the exoskeleton of the upper limbs for rehabilitation and training tasks using brain-computer interfaces.*

**Key words:** exoskeleton; actuators; brain-computer interface; motor activity training.

### References

1. He, B. Noninvasive brain-computer interfaces based on sensorimotor rhythms / B. He, B. Baxter, B. J. Edelman [et al.] // Proceedings of the IEEE. – 2015. – Vol. 103, No. 6. – P. 907–925. – DOI 10.1109/JPROC. 2015.2407272.
2. Ang, K. K. Brain-computer interface for neurorehabilitation of upper limb after stroke / K. K. Ang, C. Guan // Proceedings of the IEEE. – 2015. – Vol. 103, No. 6. – P. 944–953. – DOI 10.1109/JPROC. 2015.2415800.
3. Müller-Putz, G. R. Moregrasp: Restoration of upper limb function in individuals with high spinal cord injury by multimodal neuroprostheses for interaction in daily activities / G. R. Müller-Putz, P. Ofner, A. Schwarz [et al.] // 7<sup>th</sup> Graz Brain-Computer Interface Conference. – 2017. – Vol. 7. – P. 338–343. – DOI 10.3217/978-3-85125-533-1-62.
4. Murphy, D. P. Electroencephalogram-based brain-computer interface and lower-limb prosthesis control: A case study / D. P. Murphy, O. Bai, A. S. Gorgey [et al.] // Front. Neurol. – 2017. – Vol. 8. – P. 696. – DOI 10.3389/fneur.2017.00696.
5. Bowsher, K. Brain-computer interface devices for patients with paralysis and amputation: a meeting report / K. Bowsher, E. F. Civillico, J. Coburn [et al.] // J. Neural Eng. – 2016. – Vol. 13, No. 2. – P. 23001. – DOI 10.1088/1741-2560/13/2/023001.

# ОПРЕДЕЛЕНИЕ МОМЕНТА ВРЕМЕНИ НАЧАЛА СОВЕРШЕНИЯ ВООБРАЖАЕМОГО ДВИЖЕНИЯ ДЛЯ ЗАДАЧ РЕАБИЛИТАЦИИ

В. С. Хорев <sup>1</sup>, Н. В. Смирнов <sup>2</sup>, С. А. Куркин <sup>1, 3</sup>, А. П. Сергеев <sup>4, 5</sup>, С. Ю. Гордлеева <sup>1</sup>,  
А. Е. Храмов <sup>1, 4</sup>

<sup>1</sup> Федеральное государственное автономное образовательное учреждение высшего образования «Балтийский федеральный университет им. Иммануила Канта», Балтийский центр нейротехнологий и искусственного интеллекта, г. Калининград, Россия

<sup>2</sup> Институт разработки программного обеспечения, Университет Иннополис, г. Иннополис, Россия

<sup>3</sup> НИИ Нейронаук, Самарский государственный медицинский университет, г. Самара, Россия

<sup>4</sup> Уральский Федеральный университет им. первого Президента России Б. Н. Ельцина, г. Екатеринбург, Россия

<sup>5</sup> Институт промышленной экологии УрО РАН, г. Екатеринбург, Россия

*Работа посвящена разработке метода определения момента времени начала акта воображения движения по записям сигналов электроэнцефалограмм, которые были получены в ходе экспериментального исследования. В ходе эксперимента производилась запись электрических сигналов активности головного мозга. Сигналы, прошедшие предварительную обработку и фильтрацию, использовались для тестирования метода определения начала движения. Для демонстрации работы метода были построены зависимости усредненной мощности в альфа-диапазоне и топограммы. Полученные алгоритмы были использованы для выявления и идентификации паттернов нейронной активности, возникающих при воображении движений, а также их особенностей, по сравнению с реальной двигательной активностью и наблюдением за движениями.*

**Ключевые слова:** воображаемое движение; ээг; анализ данных; нейронная активность; эксперимент.

## 1. Введение

Способность человека представлять и моделировать новые объекты, ощущения и концепции в уме без непосредственного воздействия на чувства известна как воображение. Это сложное явление, которое трудно изучать, понимать и объяснять. Изучение активности мозга по управлению моторной активностью – сложная задача на стыке нейробиологии, медицины, нелинейной физики и инженерии. Эта проблема тесно связана с нейрореабилитацией пациентов, перенесших инсульт и страдающих двигательными и когнитивными нарушениями.

Публикации в научных журналах, посвященные экспериментальному изучению сущности воображения, появились в начале XX в. [1]. С тех пор, благодаря усилиям ученых-когнитивистов, сложилась богатая теоретическая база, которая пытается объяснить природу воображения. Важнейшим фактом, полученным недавними исследованиями воображаемой двигательной активности, является различие типов образов на основании попыток испытуемого воспроизвести их в своем уме. Такие различия существуют между воображаемым движением и моторной активностью [2].

Оценка моторного воображения имеет широкий спектр применений: разработка интерфейсов «мозг-компьютер» [3, 4], реабилитационные задачи [5, 6], прогнозирование и профилактика нейродегенеративных расстройств [7–10], исследования и диагностика психозов [11], спорт [12, 13]. Мы же рассматриваем важную проблему, связанную с исследованием возможности улучшения показателей воображаемого движения с помощью внешних воздействий на мозг в форме транскраниальной магнитной стимуляции. Это приводит к лучшему пониманию природы моторного воображения, что может помочь в будущем в разработке теорий моторного воображения и обеспечить доказанное описание физиологических процессов, лежащих в основе двигательного воображения, в качестве эталонной модели для оценки образов в будущих исследованиях.

Цель данной работы – разработка метода определения момента времени начала акта воображения движения по записям сигналов электроэнцефалограмм (ЭЭГ).

## 2. Материалы и методы

Для разработки алгоритма определения начала воображаемых движений был проведен специально разработанный эксперимент: велась запись сигналов совершения человеком движений. В экспериментах приняли участие 30 человек из числа сотрудников и студентов АНО ВО «Университет Иннополис». Отобранные участники эксперимента не имели медицинской истории черепно-мозговых травм, инсультов или неврологических заболеваний.

Во время эксперимента регистрировалась электрическая активность головного мозга в виде ЭЭГ [14]. Для этого использовалось оборудование, имеющееся в распоряжении Лаборатории нейронауки и когнитивных технологий.

Для регистрации ЭЭГ-активности использовался электроэнцефалограф actiChamp (Brain Products, Германия). Сигналы ЭЭГ были зарегистрированы для 31 канала, которые располагались неинвазивно на коже скальпа испытуемого в соответствии со схемой «10–10» (рис. 1). Заземление располагалось на месте электрода Fpz, а электрод, выступающий в роли референта, помещался за правым ухом. Для регистрации ЭЭГ использовались активные Ag/AgCl электроды ActiCA, которые помещаются в гнезда специальной шапочки EasyCAP. Кожа головы предварительно обрабатывалась абразивным гелем NuPrep, который служит для очищения и обезжиривания поверхности кожи. Затем электроды устанавливали с использованием проводящего геля SuperVisc. Данные процедуры применялись для обеспечения лучшей проводимости между кожей и электродами, что приводит к повышению общего качества регистрируемых сигналов ЭЭГ. Во время эксперимента также отслеживали значения проводимости на каждом из электродов ЭЭГ: значения меньше 25 кОм являются достаточными для правильной работы активных электродов ЭЭГ и получения сигналов хорошего качества. Частота дискретизации сигналов ЭЭГ составляет 1 000 Гц.

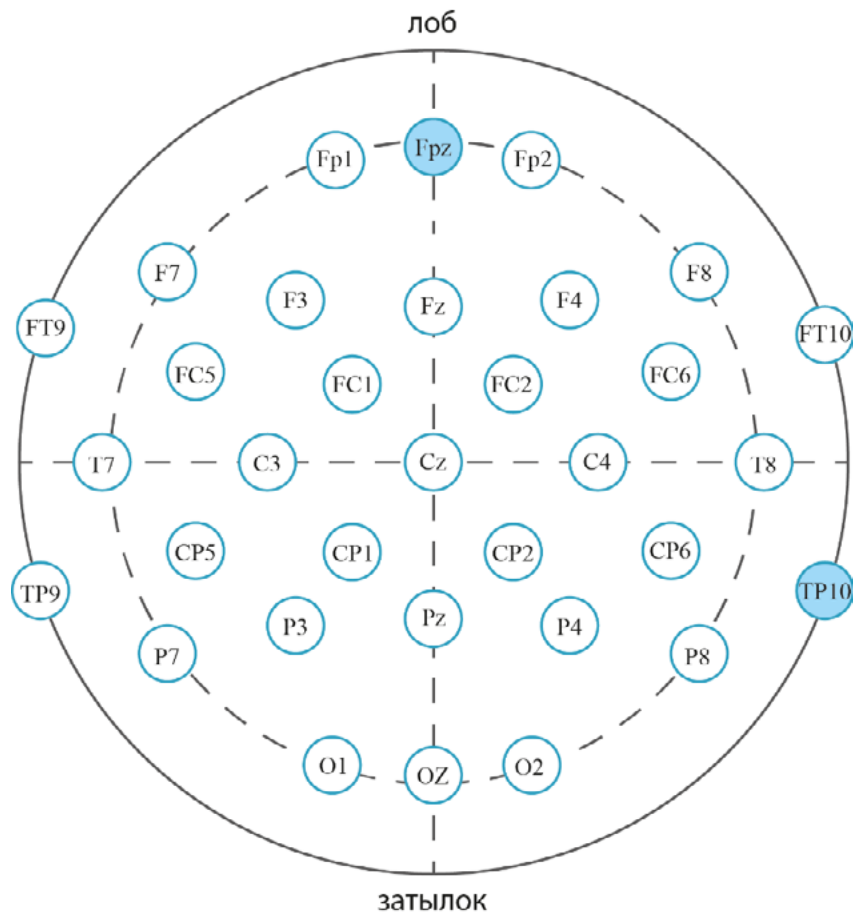


Рис. 1. Расположение электродов ЭЭГ в соответствии со стандартной схемой «10–10». Сплошным синим цветом показаны электрод заземления (Fpz) и референтный электрод (TP10)

Во всех рассматриваемых экспериментальных исследованиях принимали участие добровольцы. Все добровольцы подбирались в соответствии со следующими требованиями: правши, некурящие, не занимающиеся профессионально спортом, не принимающие медикаменты, без нейрофизиологических заболеваний и заболеваний моторно-двигательного аппарата. Перед экспериментом добровольцев просили в течение 48 часов придерживаться здорового образа жизни: не менее 8 часов сна, исключить потребление алкоголя, исключить или ограничить потребление кофеиносодержащих продуктов, избегать чрезмерных физических нагрузок. Все экспериментальные работы проводились в первой половине дня в помещении с достаточным уровнем естественного освещения. Перед экспериментом проводится инструктаж испытуемых, в ходе которого их информируют о целях и методах эксперимента, а также о связанных с его проведением возможных неудобствах. Добровольцы могли задать все интересующие вопросы и получить на них исчерпывающие ответы. Каждый испытуемый подписывал бланк информированного согласия на участие в эксперименте. Все экспериментальные работы проводились в соответствии с требованиями Хельсинкской декларации [15]. Испытуемый располагался удобно в кресле, руки лежали на подлокотниках. Перед испытуемым на расстоянии примерно 80 см от глаз был установлен монитор, на котором по ходу эксперимента предъявляются текстовые команды, информирующие о текущей фазе эксперимента или необходимости совершить определенное движение. Использовались команды следующих типов:

- выполнение реального/воображаемого движения левой/правой рукой;
- отдых между двумя последовательными выполнениями движения;



- перерыв между двумя основными сессиями в эксперименте;
- запись фоновой активности в начале и конце эксперимента.

Дизайн эксперимента схематично представлен на рис. 2. Каждый эксперимент начинался и заканчивался записью фоновой активности (180 с), во время которой испытуемому требовалось находиться в спокойном состоянии, думать о приятном, стараться не совершать никаких движений, держать глаза открытыми. Основная фаза эксперимента состояла из двух сессий. Во время сессии 1 испытуемый совершал реальные движения левой или правой рукой в соответствии с текстом команды на мониторе. Во время сессии 2 испытуемый воображал движения левой или правой рукой в соответствии с текстом команды на мониторе. Воображаемые движения аналогичны движениям, реально выполняемым в сессии 1. Каждая сессия состояла из 20 движений (триалов) – 10 реальных или воображаемых движений для каждой руки. Триалы с движениями левой и правой рукой шли в случайном порядке внутри сессии. Между сессиями 1 и 2 делался короткий перерыв (180 с).

Каждый отдельный триал состоял из базового уровня (5 с), текстовой команды с указанием требуемого типа движения (15 с) и отдыха (15 с), как показано на рис. 2. Текстовая команда находилась на мониторе в течение всех 15 с, за это время испытуемому необходимо было четко выполнить движения требуемого типа несколько раз. Интервал отдыха длился от окончания предыдущей текстовой команды до начала следующей. Интервал для базового уровня брался из отдыха после выполнения предыдущего движения.

Во время анализа данных активность мозга во время базового уровня вычиталась из активности во время выполнения движения для подавления эффектов, связанных с общим состоянием испытуемого, и получения более четко выраженной динамики активности, связанной только с выполнением движения. Выполнение движения на каждом триале (рис. 2) представляет собой многократное сгибание/разгибание всех пяти пальцев к центру ладони, напоминающее сжатие/разжатие воображаемого мяча в руке. Сгибание/разгибание пальцев выполняется со скоростью, комфортной для испытуемого.

Принимая во внимание все вышеперечисленное, было предложено следующее решение. Первым шагом проводится вейвлетный анализ в промежуточном диапазоне между альфа и бета: от 10 до 14 Гц. Анализ выполняется с помощью материнского вейвлета Морле на всем доступном временном диапазоне на каналах ЭЭГ, соответствующих моторной зоне (FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6). Далее с обоих концов (считая по времени) каждой частотно-временной поверхности исключается по 0,5 с для устранения краевых эффектов, вызванных вейвлетным преобразованием. Часть данных, соответствующих престимулу, усредняется по времени, после чего используется для процентной коррекции данных, соответствующих постстимулу. Процентная коррекция осуществляется по формуле (1):

$$\bar{p}_{ij} = \frac{p_{ij} - b_i}{b_i}, \quad (1)$$

- где  $p$  – точка вейвлетной поверхности, соответствующая постстимулу,  
 $b$  – точка, соответствующая усредненной по времени вейвлетной поверхности престимула,  
 $i, j$  – индексы по частоте и времени соответственно.

Полученные данные затем усредняются по времени. Для поиска моментов десинхронизации (начала воображаемого движения) применяется кластерный тест с перестановками на основе одновыборочного  $t$ -теста. Единицей наблюдения здесь являются испытания (триалы), каждому из которых соответствует матрица значений с размерностью (1 000, 13), где 1 000 – время (от 5 до 9 с, с частотой дискретизации 250 Гц соответственно), а 13 – каналы. Перед началом тестирования строится матрица смежности для каналов и точек по времени, которая затем используется при кластеризации. Сама процедура тестирования проводится следующим образом: для оригинальных данных рассчитываются  $t$ -значения по формуле (2):

$$t = \frac{\bar{X} - \mu}{s} \sqrt{n}, \quad (2)$$

где  $\bar{X}$  – усредненное по испытаниям значение каждой точки матрицы,  
 $s$  – стандартное отклонение этого значения,  
 $\mu$  – тестируемое значение среднего, соответствующее нулю,  
 $n$  – объем выборки.

К полученным значениям затем применяется порог, соответствующий пятипроцентному критическому уровню  $t$ -распределения. Далее оставшиеся значения кластеризуются с помощью посчитанной ранее матрицы смежности. В каждом кластере может находиться минимум два значения, при этом для каждого кластера рассчитывается так называемая кластерная статистика (в нашем случае это сумма всех  $t$ -значений, попавших в кластер), перестановочное распределение которой будет составляться далее. Само распределение составляется следующим образом: выбирается некоторое количество испытаний, после чего их значения умножаются на  $-1$  и пересчитывается кластерная статистика в существующих кластерах. Поскольку количество всех возможных перестановок растет факториально с увеличением количества испытаний, для оценки перестановочного распределения был выбран метод Монте-Карло, при котором проводятся не все возможные перестановки, а лишь случайная часть.

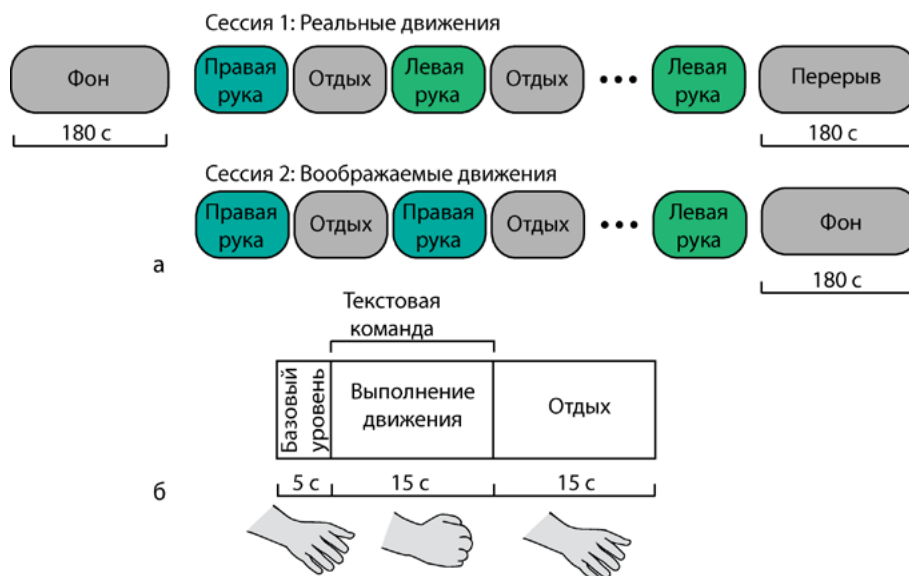


Рис. 2. Разработанный дизайн экспериментального исследования в общем виде (а) и демонстрация выполняемого движения на отдельном триале (б)

В рассматриваемом случае было выбрано количество 2 000 для каждого человека из тестовой группы. Таким образом, для каждого кластера оценивается его значимость с помощью построенного распределения. Кластер считается значимым, если его уровень значимости выше 95 %. Если же значимых кластеров не найдено, то вычисляется новое пороговое значение для  $t$ -значений с помощью списка критических уровней (5, 2,5, 1,25, 0,1 %). Полученные результаты далее используются для вычисления времени начала моторного воображения. Для этого выбирается самый ранний значимый кластер, имеющиеся  $t$ -значения в нем усредняются по значимым каналам, после чего выбирается первый локальный минимум. Время возникновения данного минимума отражает наибольшую десинхронизацию в диапазоне от 10 до 14 Гц, что отвечает требованиям активации моторной функции.

### 3. Результаты

На рис. 3 и 4 приведены результаты тестирования разработанного алгоритма для одного из добровольцев. Для выявленных кластеров справа приведены топограммы  $t$ -значений, попавших в кластер, слева – зависимости усредненной по кластеру мощности в альфа-диапазоне со стандартным отклонением (оранжевый) и  $t$ -значений, усредненных по каналам, попавших в кластер (синий); также на зависимости указан момент времени, определенный с помощью алгоритма. На представленных рисунках четко виден момент переключения  $t$ -значений для каналов, попавших в кластер, и это характеризует то, насколько хорошо работает предложенный алгоритм.

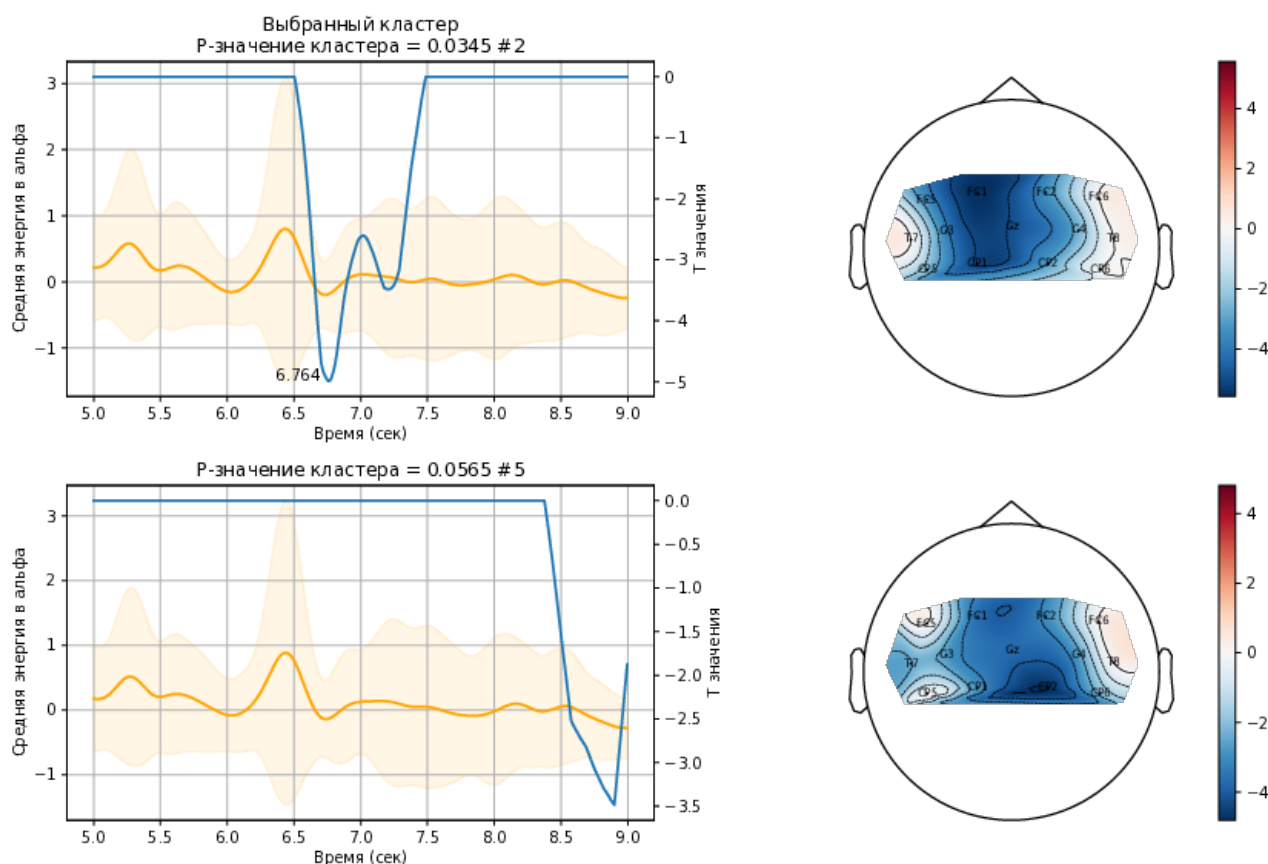


Рис. 3. Зависимости усредненной по кластеру мощности в альфа-диапазоне со стандартным отклонением (оранжевый) и  $t$ -значений, усредненных по каналам, попавших в кластер (синий), и соответствующие им топограммы

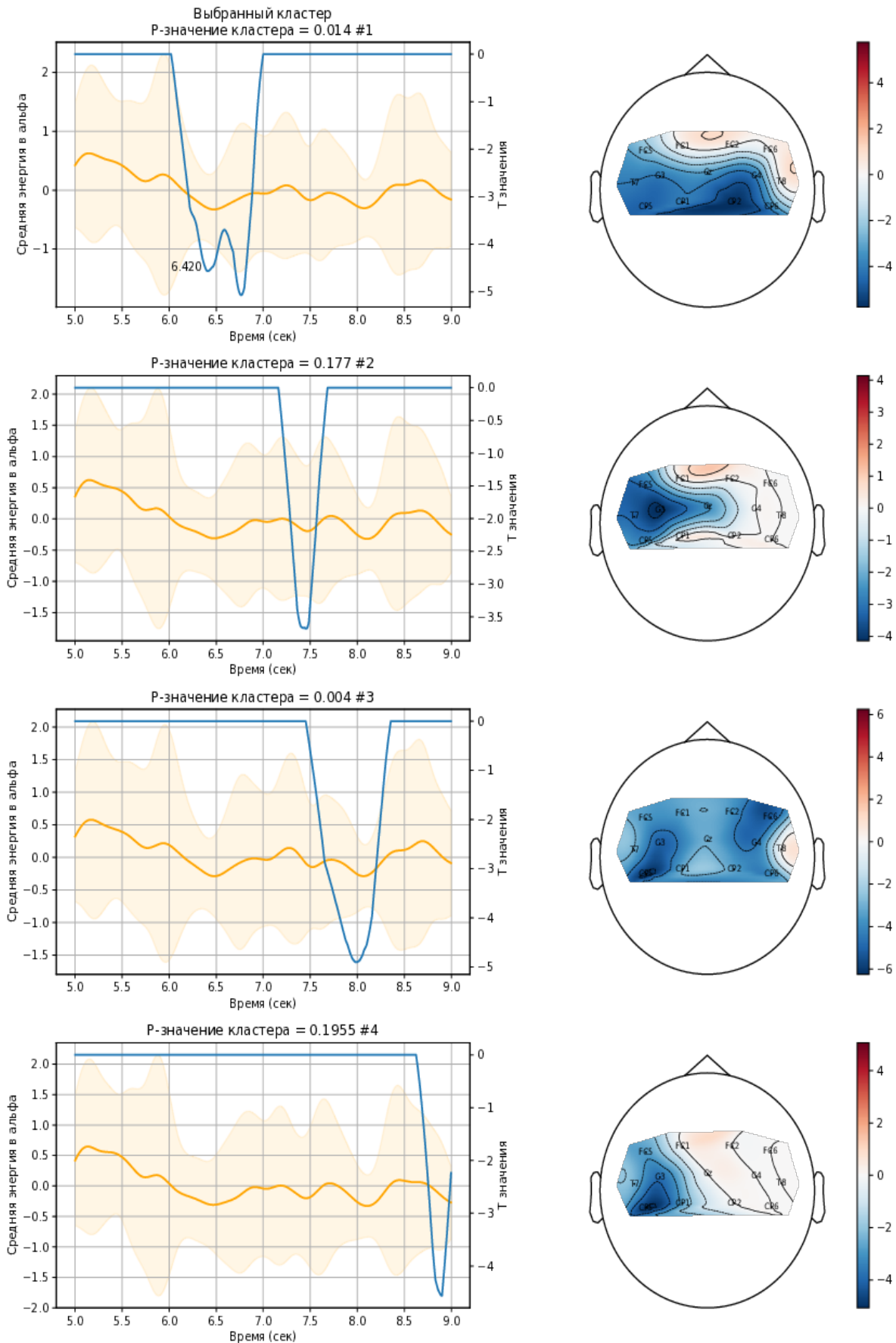


Рис. 4. Зависимости усредненной по кластеру мощности в альфа-диапазоне со стандартным отклонением (оранжевый) и  $t$ -значений, усредненных по каналам, попавших в кластер (синий), и соответствующие им топограммы

## 4. Обсуждение

Таким образом, в ходе исследования был проведен всесторонний анализ проблемы выявления и идентификации паттернов нейронной активности, возникающих при воображении движений, а также их особенностей, по сравнению с реальной двигательной активностью и наблюдением за движениями.

Записанные сигналы электрической активности прошли успешную предварительную обработку от артефактов и фильтрацию в заданном частотном диапазоне. При разработке алгоритма для определения момента времени начала акта воображения движения, который является основным в контуре управления робототехническими системами для задач реабилитации и обучения с использованием интерфейсов «мозг-компьютер» и технологий обратной связи, принималась во внимание следующая идея. Поскольку воображение является моторным, то мы должны наблюдать активацию моторной функции, что хорошо характеризуется провалом в альфа-диапазоне. Однако важно учитывать, что, в отличие от реальных движений, процессы, протекающие при моторном воображении, имеют большую вариабельность. Также следует помнить и то, что воображение относится к сложным когнитивным функциям, вследствие чего имеется тенденция проявления активности в большей степени в бета-диапазоне (от 15 до 30 Гц), нежели в альфа-диапазоне (от 8 до 12 Гц).

Помимо этого, была проведена классификация и соотнесение по областям мозга, активированным во время различных типов движений, для нахождения наиболее правдоподобных и часто наблюдаемых результатов по этому вопросу. В итоге были получены топограммы мозга, характерные для совершения воображаемого движения. Эти результаты могут помочь лучше понять природу воображаемого движения [16–18], что полезно для будущих разработок теорий, и предоставить проверенные описания физиологических процессов, лежащих в основе двигательного воображения, в качестве эталонной модели для будущей оценки воображения в новом исследовании. В свою очередь, такое исследование может быть полезно для широкой области применения, включая спорт, музыку, профилактику заболеваний и реабилитацию [7, 12].

## 5. Выводы

1. Разработан и проведен эксперимент с записью сигналов электрической активности головного мозга во время совершения воображаемых движений.
2. Проведена разработка алгоритмов определения момента времени начала акта воображения движения по записям сигналов электроэнцефалограмм.
3. Полученные алгоритмы были использованы для выявления и идентификации паттернов нейронной активности, возникающих при воображении движений, а также их особенностей, по сравнению с реальной двигательной активностью и наблюдением за движениями.

## 6. Благодарности

Исследование выполнено при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках Программы развития Уральского федерального университета им. первого Президента России Б. Н. Ельцина в соответствии с программой стратегического академического лидерства «Приоритет-2030».

## 7. Список литературы

1. *Perky, C. W.* An experimental study of imagination / C. W. Perky // *Am. J. Psychol.* – 1910. – Vol. 21. – P. 422–452. – DOI 10.2307/1413350.
2. *Sirigu, A.* Motor and Visual Imagery as Two Complementary but Neurally Dissociable Mental Processes / A. Sirigu, J. R. Duhamel // *J. Cogn. Neurosci.* – 2001. – Vol. 13. – P. 910–919. – DOI 10.1162/089892901753165827.
3. *Hramov, A. E.* Physical principles of brain-computer interfaces and their applications for rehabilitation, robotics and control of human brain states / A. E. Hramov, V. A. Maksimenko, A. N. Pisarchik // *Phys. Rep.* – 2021. – Vol. 918. – P. 1–133. – DOI 10.1016/j.physrep.2021.03.002. – EDN EBHCQC.
4. *Hramov, A. E.* Functional near-infrared spectroscopy for the classification of motor-related brain activity on the sensor-level / A. E. Hramov, V. Grubov, A. Badarin [et al.] // *Sensors.* – 2020. – Vol. 20. – P. 2362. – DOI 10.3390/s20082362. – EDN MYMKYQ.
5. *Zimmermann-Schlatter, A.* Efficacy of motor imagery in post-stroke rehabilitation: A systematic review / A. Zimmermann-Schlatter, C. Schuster, M. A. Puhan [et al.] // *J. Neuroeng. Rehabil.* – 2008. – Vol. 5. – P. 1–10. – DOI 10.1186/1743-0003-5-8.
6. *Grigorev, N. A.* A BCI-based vibrotactile neurofeedback training improves motor cortical excitability during motor imagery / N. A. Grigorev, A. O. Savosenkov, M. V. Lukoyanov [et al.] // *IEEE Trans. Neural Syst. Rehabil. Eng.* – 2021. – Vol. 29. – P. 1583–1592. – DOI 10.1109/TNSRE.2021.3102304. – EDN WDCTWF.
7. *Abidi, M.* Neural correlates of motor imagery of gait in amyotrophic lateral sclerosis / M. Abidi, G. De Marco, F. GraMO [et al.] // *J. Magn. Reson. Imaging.* – 2021. – Vol. 53. – P. 223–233. – DOI 10.1002/jmri.27335.
8. *Helmich, R. C.* Cerebral compensation during motor imagery in Parkinson's disease / R. C. Helmich, F. P. de Lange, B. R. Bloem [et al.] // *Neuropsychologia.* – 2007. – Vol. 45. – P. 2201–2215. – DOI 10.1016/j.neuropsychologia.2007.02.024.
9. *Lulé, D.* Cortical plasticity in amyotrophic lateral sclerosis: Motor imagery and function / D. Lulé, V. Diekmann, J. Kassubek [et al.] // *Neurorehabil. Neural Repair.* – 2007. – Vol. 21. – P. 518–526. – DOI 10.1177/1545968307300698.
10. *Wang, L.* Aging changes effective connectivity of motor networks during motor execution and motor imagery / L. Wang, Y. Zhang, J. Zhang [et al.] // *Front. Aging Neurosci.* 2019. – P. 312. – DOI 10.3389/fnagi.2019.00312.
11. *Danckert, J.* Attention, motor control and motor imagery in schizophrenia: Implications for the role of the parietal cortex / J. Danckert, M. Saoud, P. Maruff // *Schizophr. Res.* – 2004. – Vol. 70. – P. 241–261. – DOI 10.1016/j.schres.2003.12.007.
12. *Mizuguchi, N.* Motor imagery and sport performance / N. Mizuguchi, H. Nakata, Y. Uchida, K. Kanosue // *J. Phys. Fit. Sports Med.* – 2012. – Vol. 1. – P. 103–111. – DOI 10.7600/jpfs.1.103.
13. *Ladda, A. M.* Using motor imagery practice for improving motor performance, A review / A. M. Ladda, F. Lebon, M. Lotze // *Brain Cogn.* – 2021. – Vol. 150. – P. 105705. – DOI 10.1016/j.bandc.2021.105705 333.
14. *Buzsáki, G.* Neuronal oscillations in cortical networks / G. Buzsáki, A. Draguhn // *Science.* – 2004. – Vol. 304, No. 5679. – P. 1926–1929. – DOI 10.1126/science.1099745. – EDN HEEDPV.
15. *Rickham, P. P.* Human experimentation. Code of ethics of the world medical association. Declaration of Helsinki / P. P. Rickham // *Br. Med. J.* – 1964. – Vol. 5402, No. 2. – P. 177. – DOI 10.1136/bmj.2.5402.177.

16. *Chholak, P.* Visual and kinesthetic modes affect motor imagery classification in untrained subjects / P. Chholak, G. Niso, V. A. Maksimenko [et al.] // *Sci. Rep.* – 2019. – Vol. 9. – P. 9838. – DOI 10.1038/s41598-019-46310-9.
17. *Pisarchik, A. N.* / A. N. Pisarchik, S. A. Kurkin [et al.] // *Cybern. Phys.* – 2019. – Vol. 8. – No. 3. – P. 106–113. – DOI 10.1038/s41598-019-46310-9.
18. *Hramov, A. E.* Functional Near-Infrared Spectroscopy for the Classification of Motor-Related Brain Activity on the Sensor-Level / A. E. Hramov, V. Grubov, A. Badarin [et al.] // *Sensors.* – 2020. – Vol. 20. – P. 2362. – DOI: 10.3390/s20082362.

## Сведения об авторах:

**Хорев Владимир Сергеевич**, канд. ф.-м. н., старший научный сотрудник Балтийского центра нейротехнологий и искусственного интеллекта, Балтийский федеральный университет им. И. Канта, ул. А. Невского, 14, г. Калининград, 236041, Россия. Эл. почта: khorevvs@gmail.com.

**Смирнов Никита Владимирович**, студент АНО ВО Университет Иннополис, Университетская ул., 1, Иннополис, Респ. Татарстан, 420500, Россия. Эл. почта: n.smirnov@innopolis.university.

**Куркин Семен Андреевич**, д. ф.-м. н., доцент, 1) ведущий научный сотрудник Балтийского центра нейротехнологий и искусственного интеллекта, Балтийский федеральный университет им. И. Канта, ул. А. Невского, 14, г. Калининград, 236041, Россия; 2) главный специалист НИИ Нейронаук, Самарский государственный медицинский университет, ул. Чапаевская, 89, г. Самара, 443099, Россия. Эл. почта: kurkinsa@gmail.com.

**Сергеев Александр Петрович**, канд. ф.-м. н., заведующий лабораторией физики и экологии Института промышленной экологии УрО РАН, г. Екатеринбург, Россия. Эл. почта: aleksandrpsergeev@gmail.com.

**Гордлеева Сусанна Юрьевна**, д. ф.-м. н., 1) заведующая Научно-исследовательской лабораторией нейродинамики и когнитивных технологий, Нижегородский государственный университет им. Н. И. Лобачевского, проспект Гагарина, 23, г. Нижний Новгород, 603022, Россия; 2) старший научный сотрудник Балтийского центра нейротехнологий и искусственного интеллекта, Балтийский федеральный университет им. И. Канта, ул. А. Невского, 14, г. Калининград, 236041, Россия. Эл. почта: Gordleeva@neuro.nnov.ru.

**Храмов Александр Евгеньевич**, д. ф.-м. н., профессор, 1) главный научный сотрудник Балтийского центра нейротехнологий и искусственного интеллекта, Балтийский федеральный университет им. И. Канта, ул. А. Невского, 14, г. Калининград, 236041, Россия; 2) главный специалист НИИ Нейронаук, Самарский государственный медицинский университет, ул. Чапаевская, 89, г. Самара, 443099, Россия. Эл. почта: hramovae@gmail.com.



## DETECTION OF THE BEGINNING OF THE IMAGINARY MOVEMENT

V. S. Khorev <sup>1</sup>, N. V. Smirnov <sup>2</sup>, S. A. Kurkin <sup>1,3</sup>, A. P. Sergeev <sup>4,5</sup>,  
S. Y. Gordleeva <sup>1</sup>, A. E. Hramov <sup>1,4</sup>

<sup>1</sup> *Baltic Center for Artificial Intelligence and Neurotechnology, Immanuel Kant Baltic Federal University, Kaliningrad, Russia*

<sup>2</sup> *Institute of software development and engineering, Innopolis University, Innopolis, Russia*

<sup>3</sup> *Neurosciences Research Institute of Samara State Medical University, Samara, Russia*

<sup>4</sup> *Ural Federal University named after the first President of Russia B. N. Yeltsin, Ekaterinburg, Russia*

<sup>5</sup> *Institute of Industrial Ecology, Ural Branch of the Russian Academy of Sciences, Ekaterinburg, Russia*

*This work is devoted to the development of a method for determining the moment of the beginning of the imaginary motion act by recording electroencephalogram signals from the data obtained in the course of an experimental study of the imaginary motion acts. In the course of the experiment, electrical signals of the brain activity were recorded. The signals that were pre-processed and filtered then were used to test the method of determining the beginning of movement. To demonstrate the operation of the method, the dependences of the averaged power in the alpha-band and the topograms were plotted. The resulting algorithms were used to detect and identify the patterns of neural activity that emerge during imagery movements, as well as their features, in comparison with real motor activity and observation of movements.*

**Key words:** imaginary movement; eeg; data analysis; neural activity; experiment.

### References

1. *Perky, C. W. An experimental study of imagination / C. W. Perky // Am. J. Psychol. – 1910. – Vol. 21. – P. 422–452. – DOI 10.2307/1413350.*
2. *Sirigu, A. Motor and Visual Imagery as Two Complementary but Neurally Dissociable Mental Processes / A. Sirigu, J. R. Duhamel // J. Cogn. Neurosci. – 2001. – Vol. 13. – P. 910–919. – DOI 10.1162/089892901753165827.*
3. *Hramov, A. E. Physical principles of brain-computer interfaces and their applications for rehabilitation, robotics and control of human brain states / A. E. Hramov, V. A. Maksimenko, A. N. Pisarchik // Phys. Rep. – 2021. – Vol. 918. – P. 1–133. – DOI 10.1016/j.physrep.2021.03.002. – EDN EBHCQC.*
4. *Hramov, A. E. Functional near-infrared spectroscopy for the classification of motor-related brain activity on the sensor-level / A. E. Hramov, V. Grubov, A. Badarin [et al.] // Sensors. – 2020. – Vol. 20. – P. 2362. – DOI 10.3390/s20082362. – EDN MYMKYQ.*

5. *Zimmermann-Schlatter, A.* Efficacy of motor imagery in post-stroke rehabilitation: A systematic review / A. Zimmermann-Schlatter, C. Schuster, M. A. Puhan [et al.] // *J. Neuroeng. Rehabil.* – 2008. – Vol. 5. – P. 1–10. – DOI 10.1186/1743-0003-5-8.
6. *Grigorev, N. A.* A BCI-based vibrotactile neurofeedback training improves motor cortical excitability during motor imagery / N. A. Grigorev, A. O. Savosenkov, M. V. Lukoyanov [et al.] // *IEEE Trans. Neural Syst. Rehabil. Eng.* – 2021. – Vol. 29. – P. 1583–1592. – DOI 10.1109/TNSRE.2021.3102304. – EDN WDCTWF.
7. *Abidi, M.* Neural correlates of motor imagery of gait in amyotrophic lateral sclerosis / M. Abidi, G. De Marco, F. GraMO [et al.] // *J. Magn. Reson. Imaging.* – 2021. – Vol. 53. – P. 223–233. – DOI 10.1002/jmri.27335.
8. *Helmich, R. C.* Cerebral compensation during motor imagery in Parkinson's disease / R. C. Helmich, F. P. de Lange, B. R. Bloem [et al.] // *Neuropsychologia.* – 2007. – Vol. 45. – P. 2201–2215. – DOI 10.1016/j.neuropsychologia.2007.02.024.
9. *Lulé, D.* Cortical plasticity in amyotrophic lateral sclerosis: Motor imagery and function / D. Lulé, V. Diekmann, J. Kassubek [et al.] // *Neurorehabil. Neural Repair.* – 2007. – Vol. 21. – P. 518–526. – DOI 10.1177/1545968307300698.
10. *Wang, L.* Aging changes effective connectivity of motor networks during motor execution and motor imagery / L. Wang, Y. Zhang, J. Zhang [et al.] // *Front. Aging Neurosci.* 2019. – P. 312. – DOI 10.3389/fnagi.2019.00312.
11. *Danckert, J.* Attention, motor control and motor imagery in schizophrenia: Implications for the role of the parietal cortex / J. Danckert, M. Saoud, P. Maruff // *Schizophr. Res.* – 2004. – Vol. 70. – P. 241–261. – DOI 10.1016/j.schres.2003.12.007.
12. *Mizuguchi, N.* Motor imagery and sport performance / N. Mizuguchi, H. Nakata, Y. Uchida, K. Kanosue // *J. Phys. Fit. Sports Med.* – 2012. – Vol. 1. – P. 103–111. – DOI 10.7600/jpfs.1.103.
13. *Ladda, A. M.* Using motor imagery practice for improving motor performance, A review / A. M. Ladda, F. Lebon, M. Lotze // *Brain Cogn.* – 2021. – Vol. 150. – P. 105705. – DOI 10.1016/j.bandc.2021.105705 333.
14. *Buzsáki, G.* Neuronal oscillations in cortical networks / G. Buzsáki, A. Draguhn // *Science.* – 2004. – Vol. 304, No. 5679. – P. 1926–1929. – DOI 10.1126/science.1099745. – EDN HEEDPV.
15. *Rickham, P. P.* Human experimentation. Code of ethics of the world medical association. Declaration of Helsinki / P. P. Rickham // *Br. Med. J.* – 1964. – Vol. 5402, No. 2. – P. 177. – DOI 10.1136/bmj.2.5402.177.
16. *Chholak, P.* Visual and kinesthetic modes affect motor imagery classification in untrained subjects / P. Chholak, G. Niso, V. A. Maksimenko [et al.] // *Sci. Rep.* – 2019. – Vol. 9. – P. 9838. – DOI 10.1038/s41598-019-46310-9.
17. *Chholak, P.* Neuronal pathway and signal modulation for motor communication / P. Chholak, A. N. Pisarchik, S. A. Kurkin [et al.] // *Cybern. Phys.* – 2019. – Vol. 8. – No. 3. – P. 106–113. – DOI 10.1038/s41598-019-46310-9.
18. *Hramov, A. E.* Functional Near-Infrared Spectroscopy for the Classification of Motor-Related Brain Activity on the Sensor-Level / A. E. Hramov, V. Grubov, A. Badarin [et al.] // *Sensors.* – 2020. – Vol. 20. – P. 2362. – DOI: 10.3390/s20082362.

# ВЕРТИКАЛЬНОЕ РАСПРЕДЕЛЕНИЕ ХАРАКТЕРИСТИК ЗАПЫЛЕННОСТИ ПРИЗЕМНОГО СЛОЯ АТМОСФЕРНОГО ВОЗДУХА г. ЕКАТЕРИНБУРГА

И. Е. Субботина<sup>1</sup>, М. С. Ремезова<sup>2</sup>, А. Г. Бувич<sup>1</sup>, А. П. Сергеев<sup>1</sup>,  
Е. М. Баглаева<sup>1</sup>, А. В. Шичкин<sup>1</sup>, А. С. Буторова<sup>1</sup>, М. В. Сергеева<sup>1</sup>

<sup>1</sup> Институт промышленной экологии УрО РАН, г. Екатеринбург, Россия

<sup>2</sup> Уральский Федеральный университет им. первого Президента России

Б. Н. Ельцина, г. Екатеринбург, Россия

*Загрязнение твердыми частицами приземного слоя атмосферы урбанизированных территорий представляет собой опасность для здоровья населения. Информация о качестве воздуха является основой для принятия нормативно-правовых решений по сохранению здоровья. Было исследовано вертикальное распределение концентрации пылевых частиц, их элементного и дисперсного состава в приземном слое атмосферного воздуха г. Екатеринбурга на высоте от 0,5 до 10 м в течение 8 дней в апреле 2021 г. Массовая концентрация пыли представлена неоднородными данными с коэффициентом вариации больше 30 %, имеет слабую тенденцию уменьшаться с высотой. В количественном соотношении пылинки размером меньше РМ 2,5 составляют около 45 % общего количества частиц. Рентгенофлуоресцентным анализом было обнаружено 12 элементов, содержащихся в пылевых частицах, наиболее значительное содержание – серы, кальция и железа. Доля большинства металлов и кальция в твердых частицах уменьшается с высотой, содержание серы и мышьяка увеличивается. Содержание в частицах пыли меди, цинка и сурьмы постоянно на всех измеренных высотах.*

**Ключевые слова:** твердые частицы; вертикальное распределение; запыленность; дисперсный состав; мобильный пост.

## 1. Введение

Для оценки состояния атмосферного воздуха и его влияния на здоровье населения урбанизированных территорий необходима информация о содержании в атмосфере различных примесей и их пространственном распределении [1–10].

Наблюдения за состоянием атмосферы городов осуществляют территориальные подразделения Росгидромета и предприятия, загрязняющие атмосферный воздух селитебных территорий. Кроме регулярных наблюдений, в некоторых городах дополнительно проводятся эпизодические обследования и наблюдения. Пространственное размещение пунктов наблюдений определяется целевыми задачами, связанными с получением данных о характеристиках загрязнения наблюдаемых объектов на глобальном, региональном, территориальном и локальном уровнях [11–14].

Источники пыли в атмосфере могут иметь как природное, так и антропогенное происхождение и быть локализованы на разных территориях [15–17]. На долю частиц пыли, образовавшихся на поверхности почвы, приходится около 75 % глобальной аэрозольной нагрузки в атмосфере [18]. Химический состав атмосферной пыли включает соединения кремния, бериллия, алюминия, кадмия и других металлов, угольные частицы и сажевый аэрозоль, споры микроорганизмов и пыльцу растений, другие частицы органического происхождения. Также в результате химических реакций в атмосфере образуются вторичные неорганические соединения (нитраты, сульфаты, аммоний) [19, 20]. Для понимания механизмов образования частиц, формирования дымки и климатических воздействий важно знать особенности распределения числа пылинок по размерам. Однако такие наблюдения в мегаполисах очень ограничены. Исследования показывают, что различия в концентрациях частиц на уровне земли и на высотах порядка десятков метров значительно варьируются в зависимости от размера частиц [21–23].

При разработке моделей вертикальной структуры атмосферного аэрозоля необходимо принимать во внимание влияние подстилающей поверхности на генерацию и сток атмосферного аэрозоля, характер турбулентного вертикального массообмена, циркуляцию атмосферы, тип воздушной массы, химический состав фонового аэрозоля. Приземный слой атмосферной дымки высотой до 70 м включает пограничную зону, в которой наблюдаются значительные суточные вариации плотности аэрозоля [24]. Многообразие реализуемых в атмосфере и на поверхности земли условий приводит к перераспределению роли различных механизмов в генерации и трансформации поля атмосферного аэрозоля в зависимости от географических факторов и погодных условий. Модели распределения атмосферного аэрозоля строят с учетом географического фактора, метеоусловий, типа подстилающей поверхности, которая является как источником, так и стоком аэрозольных образований [22, 23].

Современные исследования дисперсного состава атмосферы в большинстве случаев проводят при помощи различного типа датчиков. Несмотря на высокий уровень развития вычислительной техники и измерительного оборудования, точность и воспроизводимость массовых и количественных концентраций, сообщаемых различными моделями, сильно различаются. Так, в [25, 26] показано, что данные о подсчете частиц ненадежны. Распределение числа и размера пылевых частиц, полученных на разных датчиках, не согласуются друг с другом. Таким образом, актуальными остаются измерения, основанные на прямом подсчете количества осевших на фильтрах частиц.

Институтом промышленной экологии УрО РАН создан и запатентован мобильный пост отбора проб пыли [27]. Пост состоит из насоса, питаемого аккумулятором, ряда газовых счетчиков, показывающих объем прокачанного воздуха, труб с фильтродержателями (аллонжами), ориентированными в разных направлениях. В 2016 г. с использованием мобильного поста была проведена серия измерений концентрации взвешенных частиц в приземном слое атмосферы г. Екатеринбурга на разных высотах. Была обнаружена особенность распределения концентрации по высоте – на 1 м было зафиксировано минимальное значение по сравнению со значениями концентраций на других высотах – от 0,5 до 2 м [1].

Цель настоящей работы – изучение особенностей вертикального распределения концентрации пыли, определение ее дисперсного и химического состава в приземном слое атмосферного воздуха г. Екатеринбурга.

## 2. Материалы и методы

Для исследования особенности распределения концентрации пыли на высоте 1 м, заключающейся в низком содержании пылевых частиц, мобильный пост был модифицирован: увеличили количество фильтров, на которые проводится пробоотбор, до 8, причем 5 фильтров приходятся на интервал высот 0,5–1,5 м, что позволяет исследовать распределение концентрации пыли с высокой подробностью. Фильтры располагаются на высоте 0,5; 0,75; 1; 1,25; 1,5; 2; 4 и 10 м.

Пункт наблюдений за загрязнением атмосферного воздуха размещался на открытой, проветриваемой со всех сторон площадке, с непылящим покрытием: на асфальте, твердом грунте, газоне, чтобы минимизировать искажения результатов измерений зелеными насаждениями, зданиями и другими препятствиями ветровому потоку.

Измерения запыленности атмосферного воздуха г. Екатеринбурга проводились 8 световых дней, с 9 до 20 апреля 2021 г. Пик загрязнения атмосферного воздуха обычно приходится на весенние месяцы – с начала апреля до середины мая, пока деревья не покроются листвой и газоны не зарастут травой [12].

Перед проведением эксперимента через последовательно соединенные счетчики прокачивалось  $10 \text{ м}^3$  воздуха по показаниям первого в ряду счетчика. Проверка показаний остальных счетчиков показала разницу в пределах 0,7 %, поэтому поправочные коэффициенты для расчета показателей загрязненности воздуха не вводились.

Предварительно измерялись масса и химический состав каждого чистого фильтра, определенные как среднее арифметическое по измерениям на трех разных участках фильтра. Мобильный пост работал около 9 ч в день, прокачивая в среднем  $30 \text{ м}^3$  воздуха. После этого фильтры взвешивали, определяли массу осевшей пыли как разницу между массами фильтра до и после прокачивания воздуха. Каждый день в месте проведения измерений фиксировались основные метеоданные (температура, атмосферное давление, относительная влажность, скорость и направление ветра).

На металлографическом микроскопе XJP-N100 (Wuzhou New Found Instrument Co., Ltd) делался снимок каждого фильтра с пылью, полученные фотографии (рис. 1) обрабатывались для удобства анализа. Для определения дисперсного состава подсчитанные в пяти полях зрения на каждом фильтре пылинки распределялись по 11 интервалам размеров: менее 1 мкм, 1–2, 2–3, 3–4, 4–5, 5–6, 6–7, 7–8, 8–9, 9–10 мкм и больше 10 мкм.

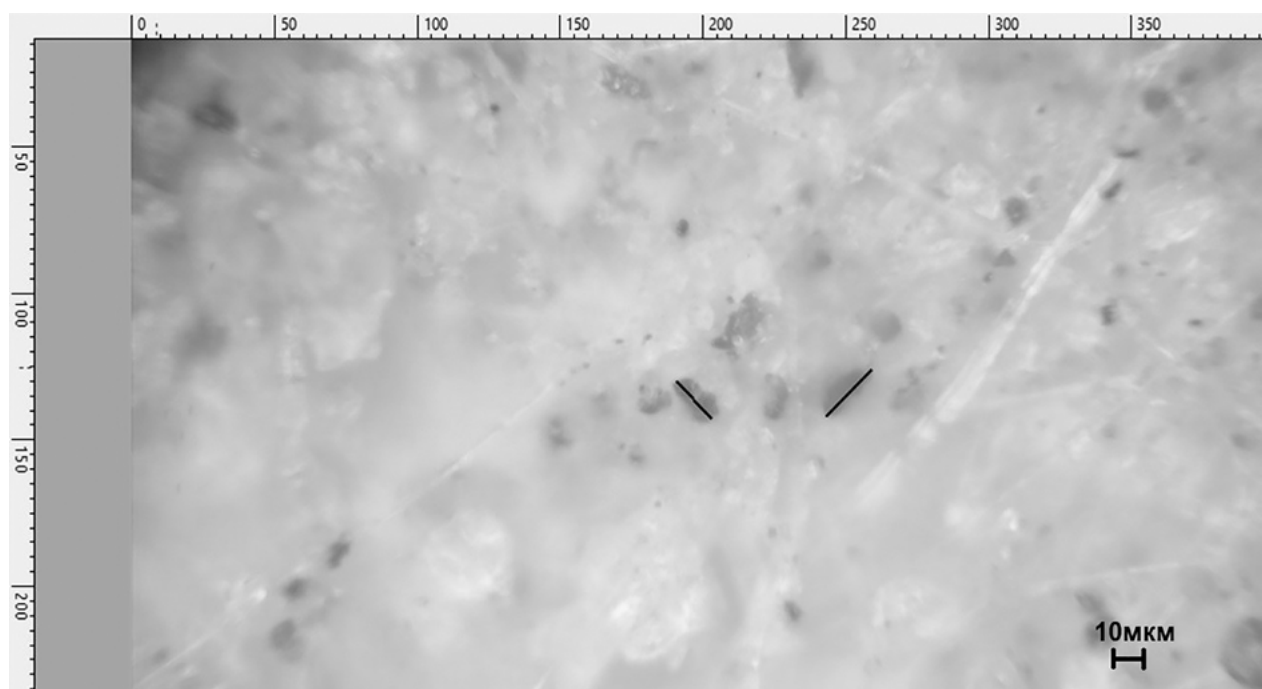


Рис. 1. Фильтр с осажденной пылью

Качественный элементный состав фильтра с осевшей на нем пылью определялся рентгенофлуоресцентным спектрометром INNOV X Systems X-5000. Технические характеристики: рентгеновская трубка с напряжением 50 кВ, током до 200 мА, разрешением 165 эВ, мощностью 10 Вт.

### 3. Результаты

Усредненные за 8 дней измерений рассчитанные концентрации пыли приведены в табл. 1. На высоте  $\leq 2$  м от поверхности земли запыленность выше, чем на высоте  $\geq 4$  м, однако в пределах 2 м никаких значимых различий в значениях концентрации пыли не обнаружено. Для оценки однородности данных определены коэффициенты вариации.

Таблица 1. Значение концентрации пыли в атмосфере г. Екатеринбурга

Высота, м	Средняя концентрация пыли, мг/м <sup>3</sup>	Среднее квадратическое отклонение, мг/м <sup>3</sup>	Коэффициент вариации
0,5	0,155	0,052	0,34
0,75	0,163	0,058	0,36
1	0,170	0,053	0,31
1,25	0,162	0,049	0,31
1,5	0,168	0,058	0,35
2	0,167	0,054	0,32
4	0,084	0,029	0,35
10	0,115	0,043	0,37

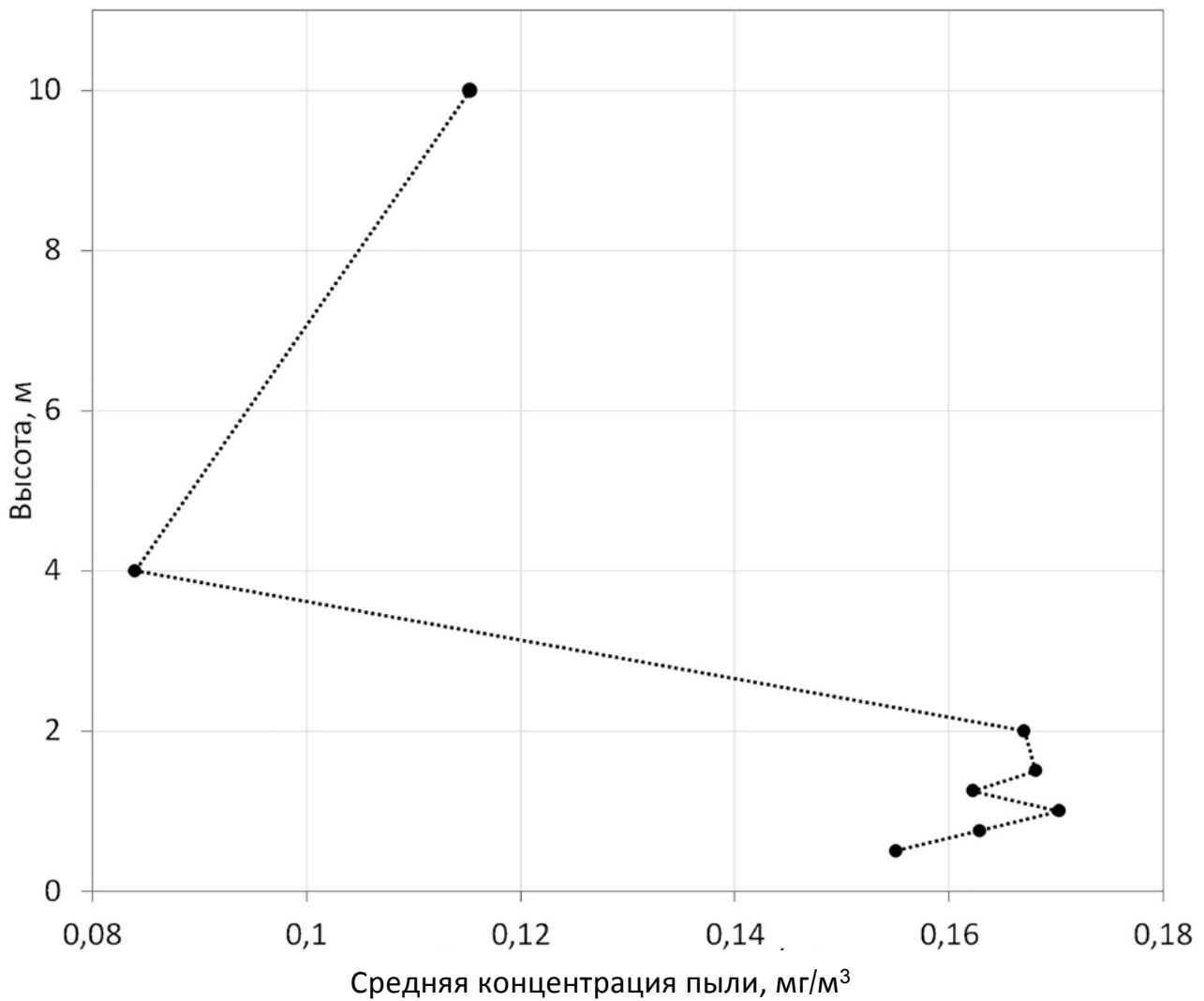


Рис. 2. Зависимость содержания пыли в атмосферном воздухе от высоты (приведена усредненная концентрация пыли за все дни измерений)

Полученные на металлографическом микроскопе фотографии обработаны, осевшие пылинки распределены по интервалам от 0 до 10 мкм с шагом 1 мкм, в последнем интервале учтены все пылинки размером больше 10 мкм. Определен дисперсный состав, который однороден для всех фильтров. На рис. 2 приведено распределение по размерам осажденных на фильтре частиц пыли для каждой высоты пробоотбора, усредненное за 8 дней. Доля мелких частиц ( $< 1$  мкм) уменьшается с увеличением высоты, исключение составляют 10 м, где их доля выросла (рис. 3). Частиц размером от 1 до 2 мкм, наоборот, тем меньше, чем ближе к поверхности земли. Пылинки остальных размеров распределены по высоте равномерно.



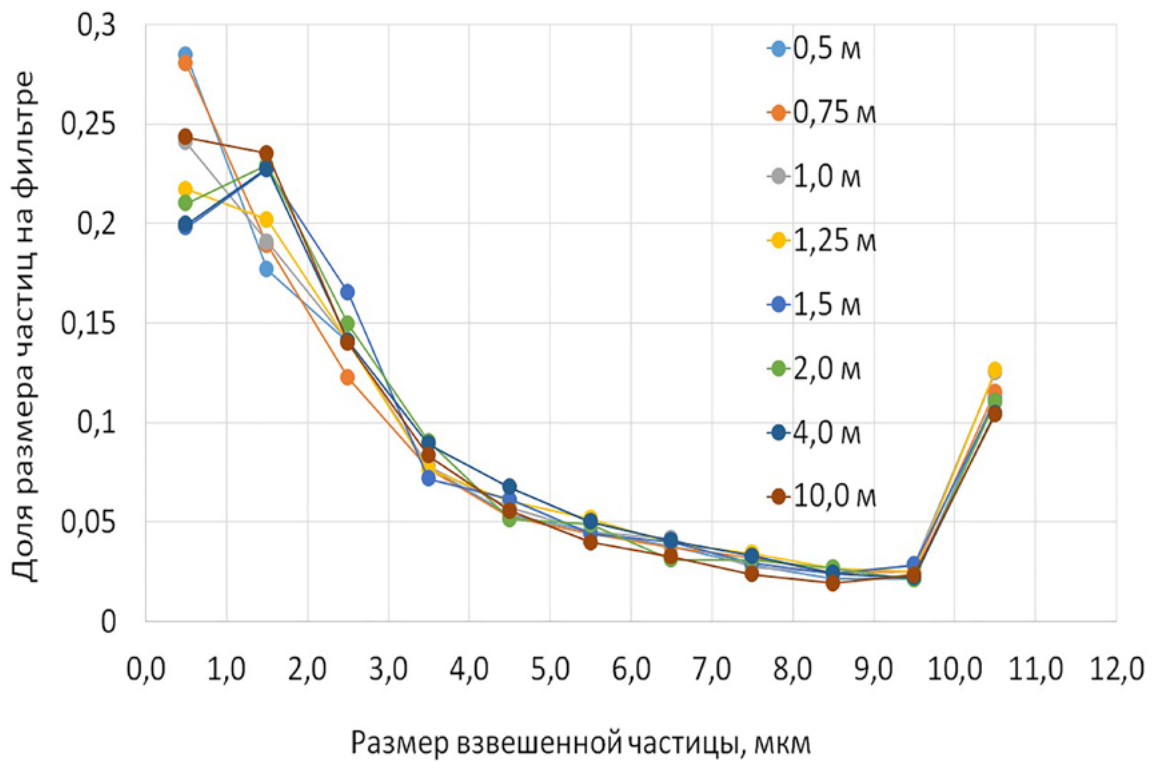


Рис. 3. Дисперсный состав пыли, осажденной на фильтре

К химическому составу пыли относили элементы, обнаруженные на прокаченном «грязном» фильтре впервые, или увеличившие свою долю по сравнению с чистым фильтром.

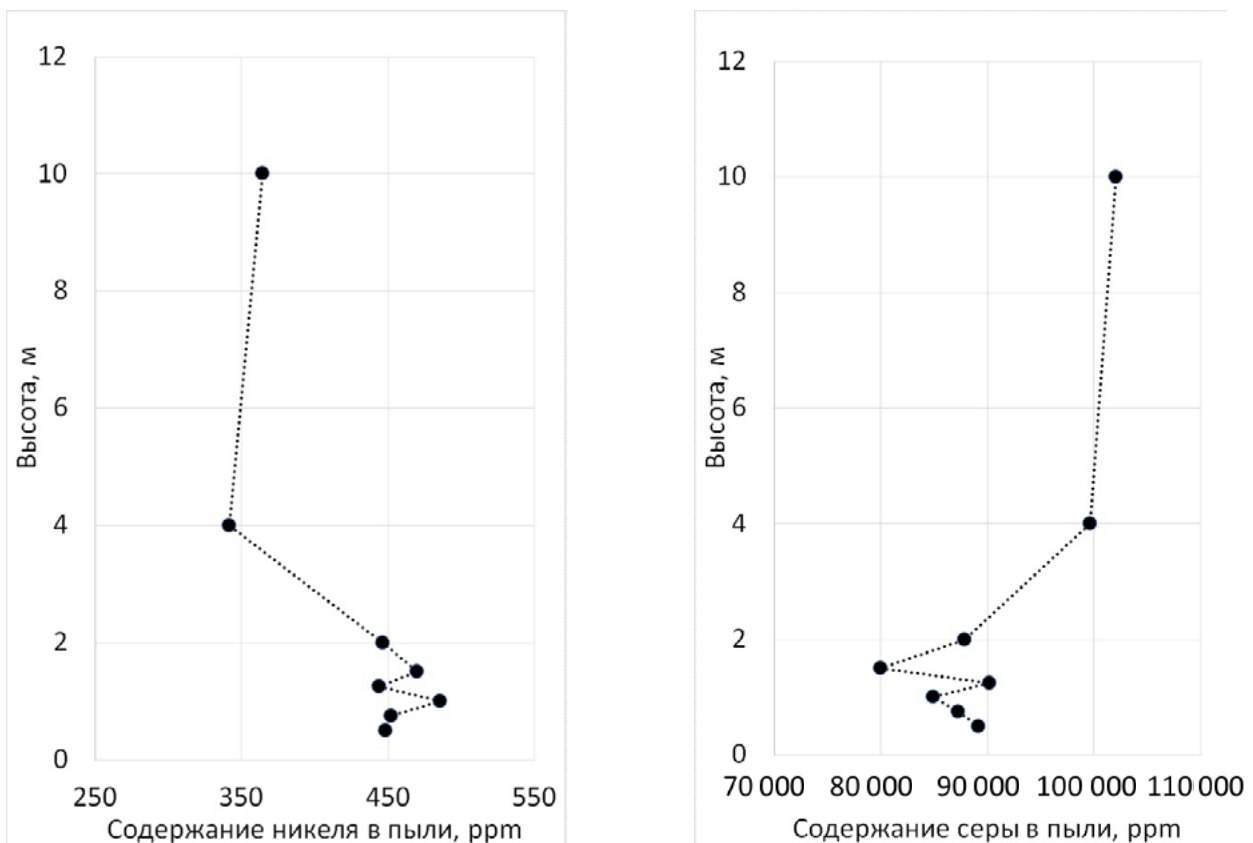


Рис. 4. Зависимость содержания химических элементов в пыли от высоты

При анализе на рентгенофлуоресцентном спектрометре были выявлены следующие элементы осевшей пыли: S (90 150 ppm), Ca (10 6140 ppm), K (2 3220 ppm), Ti (4 190 ppm), Cr (417 ppm), Mn (737 ppm), Ni (430 ppm), Cu (380 ppm), Zn (353 ppm), Sb (700 ppm), As (56 ppm), а также большое содержание Fe (83 313 ppm) (в скобках приведены усредненные по всем фильтрам значения содержания элементов). На рис. 4 приведена зависимость содержания никеля и серы в пыли от высоты.

#### 4. Обсуждение

В отличие от предыдущего исследования [1], минимальная концентрация пыли, усредненная за 8 дней измерений для исследуемой зоны дыхания человека (0,5–2 м), оказалась на высоте 0,5 м. Однако в отдельные дни измерений минимум приходился на высоты 0,75; 1,0; 1,25 м. Такие результаты могут быть связаны с особенностями воздушных потоков в приземном слое атмосферы (измерения проводились в сравнительно теплую, солнечную, безветренную погоду). Для полученных данных не было выявлено статистически значимой связи с основными метеорологическими условиями (температура, атмосферное давление, относительная влажность, скорость и направление ветра). В дни измерений был штиль (5 из 8 дней) или слабый ветер, не более 2–3 м/с. Вероятным неучтенным параметром является вертикальная составляющая скорости и силы ветра. Так, например, некоторое увеличение концентрации пыли на высоте 10 м (по сравнению с данными на 4 м) может быть связано с особенностями расположения фильтра. В отличие от других, он располагался в 3 м от крыши, на краю здания, что могло оказать влияние на распределение пыли в этой зоне.

Полученные значения коэффициентов вариации для приземной концентрации пыли, превышающие 0,3, указывают на сильные отклонения измеренных значений концентраций от среднего арифметического, что является показателем неоднородности данных. Этот результат согласуется с информацией о значительных наблюдаемых суточных изменениях запыленности в приземном слое атмосферной дымки высотой до 70 м [26].

Предельно допустимая концентрация для взвешенных частиц в атмосфере населенных мест по гигиеническим нормам ГН 2.1.6 1338-03 равна 0,5 мг/м<sup>3</sup> воздуха. Дополнением № 8 к ГН добавлено, что ПДК для пыли, состоящей из частиц диаметром менее 10 микрон, равна 0,3 мг/м<sup>3</sup>. Для пыли с частицами 2,5 мкм и менее – 0,16 мг/м<sup>3</sup>. В данном исследовании средняя концентрация пыли не превысила 0,17 мг/м<sup>3</sup>.

Отметим схожесть распределений доли элементов от высоты у металлов. Доля большинства металлов и кальция в твердых частицах уменьшается с высотой, вид распределения аналогичен рис. 4 (справа). Содержание серы и мышьяка на высоте 4 и 10 м выше, чем на высотах, не превышающих 2 м (рис. 4 (слева)). Содержание в частицах пыли меди, цинка и сурьмы постоянно на всех измеренных высотах.

#### 5. Выводы

1. Выявлены неоднородности в распределении концентрации пыли в зоне дыхания человека. Минимальное значение концентрации пыли в исследуемой зоне (0,155 мг/м<sup>3</sup>) было зафиксировано на высоте 0,5 м. Полученные значения концентраций пыли не превышают гигиенических норм (ГН 2.1.6 1338-03). Для понимания особенностей вертикального распределения пыли целесо-

образно использовать прибор для контроля вертикальной силы и скорости ветра.

2. Определены основные химические элементы, входящие в состав пыли в зоне дыхания человека: S (90 150 ppm), Ca (106 140 ppm), K (23 220 ppm), Ti (4 190 ppm), Cr (417 ppm), Mn (737 ppm), Ni (430 ppm), Cu (380 ppm), Zn (353 ppm), Sb (700 ppm), As (56 ppm) и Fe (83 313 ppm).
3. Определен дисперсный состав пыли в зоне дыхания человека. Наибольшая доля (от 0,2 до 0,27) приходится на частицы размером от 1 до 2 мкм. Доля частиц PM<sub>2,5</sub> составляет примерно 0,45.
4. Мобильный пост может быть использован в системе городского экологического мониторинга для выявления на локальных территориях уровня пылевого загрязнения атмосферного воздуха, в том числе в зоне дыхания человека.

## 6. Список литературы

1. *Baglaeva, E. M.* Particulate matter size distribution in air surface layer of Middle Ural and Arctic territories / E. M. Baglaeva, A. P. Sergeev, A. G. Buevich [et al.] // Atmospheric Pollution Research. – 2019. – № 4. – P. 1220–1226.
2. *Cachon, B. F.* Proinflammatory effects and oxidative stress within human bronchial epithelial cells exposed to atmospheric particulate matter (PM<sub>2.5</sub> and PM > 2.5) collected from Cotonou, Benin / B. F. Cachon, S. Firmin, A. Verdin [et al.] // Environmental Pollution. – 2014. – Vol. 185. – P. 340–351. <https://doi.org/10.1016/j.envpol.2013.10.026>.
3. *Hornberg, C.* Induction of sister chromatid exchanges (SCE) in human tracheal epithelial cells by the fractions PM-10 and PM-2.5 of airborne particulates / C. Hornberg, L. Maciuleviciute, N. H. Seemayer [et al.] // Toxicology Letters. – 1998. – Vol. 96–97. – P. 215–220. [https://doi.org/10.1016/S0378-4274\(98\)00075-7](https://doi.org/10.1016/S0378-4274(98)00075-7).
4. *Jia, Q.* Lung microbiome and transcriptome reveal mechanisms underlying PM<sub>2.5</sub> induced pulmonary fibrosis / Q. Jia, Q. Li, Y. Wang [et al.] // Science of The Total Environment. – 2022. – Vol. 831. – 154974. <https://doi.org/10.1016/j.scitotenv.2022.154974>.
5. *Laiman, V.* Contributions of acidic ions in secondary aerosol to PM<sub>2.5</sub> bioreactivity in an urban area / V. Laiman, Ta-Ch. Hsiao, Yu-H. Wang // Atmospheric Environment. 2022. – Vol. 275. – 119001. <https://doi.org/10.1016/j.atmosenv.2022.119001>.
6. *Lim, S. S.* A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010 / S. S. Lim, T. Vos, A. D. Flaxman [et al.] // Lancet. – 2012. – Vol. 380 (9859). P. 2224–2260. [https://doi.org/10.1016/S0140-6736\(12\)61766-8](https://doi.org/10.1016/S0140-6736(12)61766-8).
7. *Selokar, A.* PM<sub>2,5</sub> particulate matter and its effects in Delhi/NCR / A. Selokar, B. Ramachandran, K. N. Elangovan [et al.] // Materials Today: Proceedings. – 2020. – Vol. 33(7). – P. 4566–4572. <https://doi.org/10.1016/j.matpr.2020.08.187>.
8. *Wu, H.* Short-term effects of exposure to ambient PM<sub>1</sub>, PM<sub>2.5</sub>, and PM<sub>10</sub> on ischemic and hemorrhagic stroke incidence in Shandong Province, China / H. Wu, B. Zhang, J. Wei [et al.] // Environmental Research. – 2022. – Vol. 212. – Part C. – 113350. <https://doi.org/10.1016/j.envres.2022.113350>.
9. *Xu, C.* Overlooked Significant Impact of Trace Metals on the Bacterial Community of PM<sub>2.5</sub> in High-Time Resolution / C. Xu, H. Chen, C. Zhu [et al.] // J. of Geophysical Research: Atmospheres. 2021. – Vol. 126 (23). – e2021JD035408. <https://doi.org/10.1029/2021JD035408>.

10. Yang, M. Fine and ultrafine airborne PM influence inflammation response of young adults and toxicological responses in vitro / M. Yang, P. Jalava, H. Hakkarainen [et al.] // Science of The Total Environment. – 2022. – Vol. 836. – 155618. <https://doi.org/10.1016/j.scitotenv.2022.155618>.
11. Приказ Министерства природных ресурсов и экологии РФ от 30 июля 2020 г. N 524 «Об утверждении требований к проведению наблюдений за состоянием окружающей среды, ее загрязнением» (вступает в силу с 1 января 2021 г. и действует до 1 января 2027 г.).
12. Обзор состояния и загрязнения окружающей среды в Российской Федерации за 2019 год. Федеральная служба по гидрометеорологии и мониторингу окружающей среды – Москва, 2020. <http://downloads.igce.ru/publications/reviews/review2019.pdf>.
13. РД 52. 04. 186-89. Руководящий документ. Руководство по контролю загрязнения атмосферы. Введ. 1991-07-01. – Москва, 1991. – 693 с. [https://fcgie.ru/download/osnovnye\\_dokumenty\\_po\\_vedeniyu\\_sgm/186\\_89.pdf](https://fcgie.ru/download/osnovnye_dokumenty_po_vedeniyu_sgm/186_89.pdf).
14. Ключев, Н. Н. Качество атмосферного воздуха российских городов в 1991–2016 гг. / Н. Н. Ключев // Известия Российской академии наук. Серия географическая. – 2019. – № 1. – С. 14–23. <https://doi.org/10.31857/S2587-55662019114-23>.
15. Brey, S. J. Estimating the Spread in Future Fine Dust Concentrations in the Southwest United States / S. J. Brey, J. R. Pierce, E. A. Barnes [et al.] // J. of Geophysical Research: Atmospheres. – 2020. – Vol. 125 (21). e2019JD031735. <https://doi.org/10.1029/2019JD031735>.
16. Lee, G. Roles of meteorological factors in inter-regional variations of fine and coarse PM concentrations over the Republic of Korea / G. Lee, Y. G. Lee, E. Jeong [et al.] // Atmospheric Environment. 2021. – Vol. 264. – 118706. <https://doi.org/10.1016/j.atmosenv.2021.118706>.
17. Maring, H. Mineral dust aerosol size distribution change during atmospheric transport / H. Maring, D. L. Savoie, M. A. Izaguirre [et al.] // J. of Geophysical Research. 2003. Vol. 108. No. D19. 8592. – doi:10.1029/2002JD002536.
18. Романовская, А. Ю. Аэрозольная пыль почвенного происхождения в атмосфере: источники, количество, свойства (обзор) / А. Ю. Романовская, И. Ю. Савин // Бюллетень Почвенного института имени В. В. Докучаева. – 2021. – Вып. 109. – С. 36–95. DOI: 10.19047/0136-1694-2021-109-36-95.
19. Заиков, Г. Е. Кислотные дожди и окружающая среда / Г. Е. Заиков, С. А. Маслов, В. П. Рубайло [и др]. – М. : Химия, 1991. – 139 с.
20. Исидоров, В. А. Экологическая химия: учеб. пособие для вузов / В. А. Исидоров. – СПб. : Химиздат, 2001. 301 с.
21. Du, W. Insights into vertical differences of particle number size distributions in winter in Beijing, China / W. Du, W. Wang, R. Liu [et al.] // Science of The Total Environment. – 2022. – Vol. 802. – 149695. <https://doi.org/10.1016/j.scitotenv.2021.149695>.
22. Liao, T. Seasonal distribution and vertical structure of different types of aerosols in southwest China observed from CALIOP / T. Liao, K. Gui, Y. Li [et al.] // Atmospheric Environment. – 2021. – Vol. 246. – 118145. <https://doi.org/10.1016/j.atmosenv.2020.118145>.
23. Dubey, R. Evaluation of vertical and horizontal distribution of particulate matter near an urban roadway using an unmanned aerial vehicle / R. Dubey, A. K. Patra, J. Joshi [et al.] // Science of The Total Environment. – 2022. – Vol. 836. – 155600. <https://doi.org/10.1016/j.scitotenv.2022.155600>.
24. Кондратьев, К. Я. Атмосферный аэрозоль / К. Я. Кондратьев, Н. И. Москаленко, Д. В. Поздняков. – Л. : Гидрометеоиздат, 1983. – 224 с.

25. *Dubey, R.* Evaluation of low-cost particulate matter sensors OPC N2 and PM Nova for aerosol monitoring / R. Dubey, A. K. Patra, J. Joshi [et al.] // Atmospheric Pollution Research. – 2022. – Vol. 13(3). 101335. <https://doi.org/10.1016/j.apr.2022.101335>.
26. *Tryner, J.* Effects of aerosol type and simulated aging on performance of low-cost PM sensors / J. Tryner, J. Mehaffy, D. Miller-Lionberg [et al.] // J. of Aerosol Science. – 2020. – Vol. 150. – 105654. <https://doi.org/10.1016/j.jaerosci.2020.105654>.
27. *Баглаева Е. М.* Мобильный пост отбора проб пыли приземного слоя атмосферного воздуха со стратификацией по высоте / Е. М. Баглаева, А. Г. Буевич, И. Е. Субботина [и др.] // ЭСИП. – 2017. – № 7. – С. 23–32.

### **Сведения об авторах:**

**Субботина Ирина Евгеньевна**, канд. ф.-м. н., научный сотрудник Института промышленной экологии УрО РАН, г. Екатеринбург, Россия. Эл. почта: [iesub@mail.ru](mailto:iesub@mail.ru).

**Ремезова Мария Сергеевна**, студентка, ФГАОУ ВО «Уральский федеральный университет им. первого Президента России Б. Н. Ельцина», физико-технологический институт, г. Екатеринбург, Россия.

**Буевич Александр Геннадьевич**, научный сотрудник Института промышленной экологии УрО РАН, г. Екатеринбург, Россия.

**Сергеев Александр Петрович**, канд. ф.-м. н., ведущий научный сотрудник Института промышленной экологии УрО РАН, г. Екатеринбург, Россия.

**Баглаева Елена Михайловна**, канд. ф.-м. н., старший научный сотрудник Института промышленной экологии УрО РАН, г. Екатеринбург, Россия.

**Шичкин Андрей Васильевич**, научный сотрудник Института промышленной экологии УрО РАН, г. Екатеринбург, Россия.

**Буторова Анастасия Сергеевна**, инженер-исследователь Института промышленной экологии УрО РАН, г. Екатеринбург, Россия.

**Сергеева Марина Викторовна**, научный сотрудник Института промышленной экологии УрО РАН, г. Екатеринбург, Россия.

## VERTICAL DISTRIBUTION OF DUST CHARACTERISTICS OF THE ATMOSPHERIC SURFACE LAYER IN EKATERINBURG

I. E. Subbotina <sup>1</sup>, M. S. Remezova <sup>2</sup>, A. G. Buevich <sup>1</sup>, A. P. Sergeev <sup>1</sup>, E. M. Baglaeva <sup>1</sup>,  
A. V. Shichkin <sup>1</sup>, A. S. Butorova <sup>1</sup>, M. V. Sergeeva <sup>1</sup>

<sup>1</sup> Institute of Industrial Ecology UB RAS, Ekaterinburg, Russia

<sup>2</sup> Ural Federal University named after the first President of Russia B. N. Yeltsin,  
Ekaterinburg, Russia

*Pollution of the surface layer of the atmosphere with particulate matter in urban areas is a danger to public health. Air quality information is the basis for health policy making. The paper investigates the vertical distribution of the concentration of dust particles, their elemental and dispersion composition in the surface layer of atmospheric air in Yekaterinburg at a height of 0,5 m to 10 m during 8 days of April 2021. The mass concentration of dust is represented by heterogeneous data with a variation coefficient of more than 30 %, it has a weak tendency to decrease with height. In quantitative terms, dust particles PM<sub>2.5</sub> make up about 45 % of the total number of particles. X-ray fluorescence analysis revealed 12 elements contained in dust particles, the most significant content being sulfur, calcium and ferrum. The proportion of the most metals and calcium in particulate matter decreases with height, while the content of sulfur and arsenicum increases. The content of cuprum, zincum and stibium in particulate matter is constant at all measured heights.*

**Key words:** particulate matter, vertical distribution, dust content, dispersed composition, mobile post.

### References

1. Baglaeva, E. M. Particulate matter size distribution in air surface layer of Middle Ural and Arctic territories / E. M. Baglaeva, A. P. Sergeev, A. G. Buevich [et al.] // Atmospheric Pollution Research. – 2019. – № 4. – P. 1220–1226.
2. Cachon, B. F. Proinflammatory effects and oxidative stress within human bronchial epithelial cells exposed to atmospheric particulate matter (PM<sub>2.5</sub> and PM > 2.5) collected from Cotonou, Benin / B. F. Cachon, S. Firmin, A. Verdin [et al.] // Environmental Pollution. – 2014. – Vol. 185. – P. 340–351. <https://doi.org/10.1016/j.envpol.2013.10.026>.
3. Hornberg, C. Induction of sister chromatid exchanges (SCE) in human tracheal epithelial cells by the fractions PM-10 and PM-2.5 of airborne particulates / C. Hornberg, L. Maciuleviciute, N. H. Seemayer [et al.] // Toxicology Letters. – 1998. – Vol. 96–97. – P. 215–220. [https://doi.org/10.1016/S0378-4274\(98\)00075-7](https://doi.org/10.1016/S0378-4274(98)00075-7).
4. Jia, Q. Lung microbiome and transcriptome reveal mechanisms underlying PM<sub>2.5</sub> induced pulmonary fibrosis / Q. Jia, Q. Li, Y. Wang [et al.] // Science of The Total Environment. – 2022. – Vol. 831. – 154974. <https://doi.org/10.1016/j.scitotenv.2022.154974>.

5. *Laiman, V.* Contributions of acidic ions in secondary aerosol to PM<sub>2.5</sub> bioreactivity in an urban area / V. Laiman, Ta-Ch. Hsiao, Yu-H. Wang // *Atmospheric Environment*. 2022. – Vol. 275. – 119001. <https://doi.org/10.1016/j.atmosenv.2022.119001>.
6. *Lim, S. S.* A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010 / S. S. Lim, T. Vos, A. D. Flaxman, et al // *Lancet*. – 2012. – Vol. 380 (9859). P. 2224–2260. [https://doi.org/10.1016/S0140-6736\(12\)61766-8](https://doi.org/10.1016/S0140-6736(12)61766-8).
7. *Selokar, A.* PM 2,5 particulate matter and its effects in Delhi/NCR / A. Selokar, B. Ramachandran, K. N. Elangovan [et al.] // *Materials Today: Proceedings*. – 2020. – Vol. 33(7). – P. 4566–4572. <https://doi.org/10.1016/j.matpr.2020.08.187>.
8. *Wu, H.* Short-term effects of exposure to ambient PM<sub>1</sub>, PM<sub>2.5</sub>, and PM<sub>10</sub> on ischemic and hemorrhagic stroke incidence in Shandong Province, China / H. Wu, B. Zhang, J. Wei [et al.] // *Environmental Research*. – 2022. – Vol. 212. – Part C. – 113350. <https://doi.org/10.1016/j.envres.2022.113350>.
9. *Xu, C.* Overlooked Significant Impact of Trace Metals on the Bacterial Community of PM<sub>2.5</sub> in High-Time Resolution / C. Xu, H. Chen, C. Zhu [et al.] // *J. of Geophysical Research: Atmospheres*. 2021. – Vol. 126 (23). – e2021JD035408. <https://doi.org/10.1029/2021JD035408>.
10. *Yang, M.* Fine and ultrafine airborne PM influence inflammation response of young adults and toxicological responses in vitro / M. Yang, P. Jalava, H. Hakkarainen [et al.] // *Science of The Total Environment*. – 2022. – Vol. 836. – 155618. <https://doi.org/10.1016/j.scitotenv.2022.155618>.
11. Order of the Ministry of Natural Resources and Ecology of the Russian Federation from July 30, 2020 N 524 «On approval of the requirements for monitoring the state of the environment, its pollution» (Entering into force on January 1, 2021 and valid until January 1, 2027). (In Russian).
12. Overview of the state and pollution of the environment in the Russian Federation for 2019. Federal Service for Hydrometeorology and Environmental Monitoring. – Moscow. – 2020. – <http://downloads.igce.ru/publications/reviews/review2019.pdf> (In Russian).
13. RD 52. 04. 186-89. Guidance document. Air Pollution Control Guide. – Input. 1991-07-01. – Moscow. – 1991. – 693 p. – [https://fcgie.ru/download/osnovnye\\_dokumenty\\_po\\_vedeniyu\\_sgm/186\\_89.pdf](https://fcgie.ru/download/osnovnye_dokumenty_po_vedeniyu_sgm/186_89.pdf) (In Russian).
14. *Klyuyev, N. N.* Atmospheric air quality in Russian cities in 1991–2016 / N. N. Klyuev // *Proceedings of the Russian Academy of Sciences*. – Geographic series. – 2019. – No. 1. – P. 14–23. – <https://doi.org/10.31857/S2587-55662019114-23> (In Russian).
15. *Brey, S. J.* Estimating the Spread in Future Fine Dust Concentrations in the Southwest United States / S. J. Brey, J. R. Pierce, E. A. Barnes [et al.] // *J. of Geophysical Research: Atmospheres*. – 2020. – Vol. 125 (21). e2019JD031735. <https://doi.org/10.1029/2019JD031735>.
16. *Lee, G.* Roles of meteorological factors in inter-regional variations of fine and coarse PM concentrations over the Republic of Korea / G. Lee, Y. G. Lee, E. Jeong [et al.] // *Atmospheric Environment*. 2021. – Vol. 264. – 118706. <https://doi.org/10.1016/j.atmosenv.2021.118706>.
17. *Maring, H.* Mineral dust aerosol size distribution change during atmospheric transport / H. Maring, D. L. Savoie, M. A. Izaguirre [et al.] // *J. of Geophysical Research*. 2003. Vol. 108. No. D19. 8592. doi:10.1029/2002JD002536.

18. *Romanovskaya, A. Yu.* Aerosol dust of soil origin in the atmosphere: sources, quantity, properties (review) / A. Y. Romanovskaya, I. Yu. Savin // Bulletin of the Soil Institute named after V. V. Dokuchaev. – 2021. – Issue. 109. P. 36–95. DOI: 10.19047/0136-1694-2021-109-36-95 (In Russian).
19. *Zaikov, G. E.* Acid rains and the environment / G. E. Zaikov, S. A. Maslov, V. P. Rubailo. – Moscow : Chemistry, – 1991. – 139 p. (In Russian).
20. *Isidorov, V. A.* Ecological chemistry: Textbook for universities. / V. A. Isidorov // St. Petersburg : Himizdat. – 2001. – 301 p. (In Russian).
21. *Du, W.* Insights into vertical differences of particle number size distributions in winter in Beijing, China / W. Du, W. Wang, R. Liu [et al.] // Science of The Total Environment. – 2022. – Vol. 802. – 149695. <https://doi.org/10.1016/j.scitotenv.2021.149695>.
22. *Liao, T.* Seasonal distribution and vertical structure of different types of aerosols in southwest China observed from CALIOP / T. Liao, K. Gui, Y. Li [et al.] // Atmospheric Environment. – 2021. – Vol. 246. – 118145. <https://doi.org/10.1016/j.atmosenv.2020.118145>.
23. *Dubey, R.* Evaluation of vertical and horizontal distribution of particulate matter near an urban roadway using an unmanned aerial vehicle / R. Dubey, A. K. Patra, J. Joshi [et al.] // Science of The Total Environment. – 2022. – Vol. 836. – 155600. <https://doi.org/10.1016/j.scitotenv.2022.155600>.
24. *Kondratiev, K. Y.* Atmospheric aerosol. / K. Y. Kondratiev, N. I. Moskalenko, D. V. Pozdnyakov. – Leningrad : Gidrometeoizdat, 1983. – 224 p. (In Russian).
25. *Dubey, R.* Evaluation of low-cost particulate matter sensors OPC N2 and PM Nova for aerosol monitoring / R. Dubey, A. K. Patra, J. Joshi [et al.] // Atmospheric Pollution Research. – 2022. – Vol. 13(3). 101335. <https://doi.org/10.1016/j.apr.2022.101335>.
26. *Tryner, J.* Effects of aerosol type and simulated aging on performance of low-cost PM sensors / J. Tryner, J. Mehaffy, D. Miller-Lionberg [et al.] // J. of Aerosol Science. – 2020. – Vol. 150. – 105654. <https://doi.org/10.1016/j.jaerosci.2020.105654>.
27. *Baglaeva, E. M.* Mobile station for dust sampling of the surface layer of atmospheric air with height stratification / E. M. Baglaeva, A. G. Buevich, I. E. Subbotina [et al.] // ESIP. – 2017. – No. 7. – P. 23–32. (In Russian).



# СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ РАЗДЕЛЕНИЯ ДАННЫХ ДЛЯ ОБУЧЕНИЯ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ ПРИ МОДЕЛИРОВАНИИ ПРОСТРАНСТВЕННОГО РАСПРЕДЕЛЕНИЯ ТЯЖЕЛЫХ МЕТАЛЛОВ В ВЕРХНЕМ СЛОЕ ПОЧВЫ

Е. М. Баглаева, А. П. Сергеев, А. В. Шичкин, А. Г. Буевич,  
И. Е. Субботина, А. С. Буторова

*Институт промышленной экологии УрО РАН, г. Екатеринбург, Россия*

*Представлен сравнительный анализ пяти алгоритмов выбора обучающего подмножества для искусственных нейронных сетей, интерполирующих пространственное распределение признака по данным экологических скринингов компонентов окружающей природной среды. В качестве исходных данных использовали содержание хрома в верхнем слое почвы на селитебных территориях г. Ноябрьска (ЯНАО, Россия). Пространственные распределения содержания элементов в верхнем слое почвы интерполировались многослойным персептроном (МЛП). Статистики среднеквадратической ошибки (RMSE) рассчитывались как показатели точности для каждого алгоритма. Меньшую ошибку демонстрирует основанный на пространственном кватеринге исходных данных алгоритм Space&quot;quartile quote. Принимая во внимание информацию о статистике ошибок алгоритмов разделения исходного набора, можно повысить точность модели.*

**Ключевые слова:** МЛП; разделение исходных данных; пространственное распределение; моделирование среды; репрезентативность.

## 1. Введение

Оценка и прогнозирование состояния компонентов окружающей природной среды является важнейшей составляющей оценки антропогенного воздействия и влияния на здоровье человека [1, 2]. Верхний слой почвы – наиболее объективный и устойчивый индикатор техногенного загрязнения почвы и экосистемы [3, 4]. Почвенная съемка обычно используется для изучения характера и характеристик окружающей среды [5, 6]. Каждый образец почвы содержит информацию о локальных пространственно-временных свойствах участка отбора проб [7], которые зависят от местных климатических, гидрогеологических и геологических условий, нерегулярных источников загрязнения и других постоянно меняющихся факторов. Для объективной оценки почвенных характеристик необходимо выбрать места отбора проб и взять как можно больше проб [8]. Предварительная оценка репрезентативности

точек отбора проб может снизить количество проб и получить большой объем информации об изучаемой территории [9, 10].

Почвенная съемка используется для картографирования пространственного распределения различных характеристик почвы, например, гранулометрического состава, объемной плотности, содержания химических реагентов и других [11]. Данные съемки часто сочетаются с методологией искусственной нейронной сети (ИНС) для реконструкции пространственного распределения признака [12, 13]. Почвенные данные представляют собой результаты обследования в нерегулярных точках отбора проб и используются в качестве исходных для обучения и тестирования модели ИНС. Существуют различные методы построения наилучшего обучающего набора [14–16]. Обычно ошибка ИНС сводится к минимуму. Исследование деления данных на обучающее и тестовое подмножество может привести к созданию наилучшего набора обучающих данных для оценки эффективности различных точек выборки для решения конкретной задачи [17, 18]. Разделение данных является важным фактором при построении ИНС, метод перекрестной проверки часто используется для оценки производительности моделей ИНС [19, 20]. Методология разделения выборочных данных может оказать значительное влияние на качество обучающих, тестовых и проверочных подмножеств для ИНС. Неправильное разделение данных может привести к высокой дисперсии и систематической ошибке в полученных результатах, а также к ложному отчету о прогностической эффективности модели [21, 22]. Способ разделения данных не зависит от типа модели ИНС, для проверки возможностей метода деления удобно использовать наиболее простую для понимания модель, например, многослойный персептрон (МЛП). Нейронная сеть МЛП обучена прогнозировать закономерности пространственно-временной изменчивости для коротких периодов [23, 24]. Таким образом, оценка репрезентативности точек отбора проб имеет большое практическое значение для построения оптимальной схемы пробоотбора при экологическом мониторинге, с другой стороны – для построения наилучшего обучающего множества для ИНС.

Методы отбора наилучшей обучающей выборки позволяют оценить репрезентативность расположения наиболее информативных точек выборки. В нашей статье мы сравниваем пять полученных (с использованием различных подходов к разделению данных) способов построения обучающего подмножества для моделирования пространственного распределения тяжелых металлов в верхнем слое почвы.

## **2. Материалы и методы**

### **2.1. География отбора проб**

Почвенная съемка была организована в селитебной зоне г. Ноябрьска (N63.2°, E75.5°), расположенного у южных границ Ямало-Ненецкого автономного округа, Россия (рис. 1). Основные предприятия города связаны с добычей углеводородных ресурсов. Климат резко континентальный. Город находится в природной зоне тайги. Преобладающий тип почвы – таежные подзолы [25].

### **2.2. Отбор проб почвы, подготовка образцов и химический анализ**

Получение исходных данных состояло из трех этапов: отбор проб почвы, подготовка образцов и химический анализ. Процедура отбора проб почвы подробно описана [10]. На рис. 1 показана подробная схема пространственного отбора проб и отбор проб почвы на месте. Каждый образец был упакован в двойной полиэтилено-

вый пакет с идентификатором образца. Идентификатор был отмечен районом, географическими координатами ( $N6X.XXX^\circ$ ,  $E7X.XXX^\circ$ ) и количеством проб почвы. Упакованные образцы были доставлены в лабораторию.

Подготовка почвенных образцов включала четыре этапа (рис. 2): подсушивание на воздухе, просеивание, квартование и измельчение в соответствии с действующими стандартными требованиями, подробно описанными в [10]. Первый этап представлял собой сушку на воздухе при температуре  $20^\circ\text{C}$  и давлении 101 МПа. Вторым этапом было просеивание через сито диаметром 1 мм для удаления крупных неорганических и органических остатков. Квартирование представляло собой перемешивание и гомогенизацию для получения 20 г однородного субобразца. Четвертый – измельчение на зерно диаметром 0,74 мм для химического анализа. Химический анализ включает ( $\text{HNO}_3 + \text{HF} + \text{HClO}_4$ ) кислотное разложение и масс-спектрометрию с индуктивно связанной плазмой (Perkin Elmer, ELAN 9000). Измерено содержание в почве Si, K, Ca, V, Cr, Mn, Ni, Cu, Zn. В дальнейшем моделировали содержание хрома.

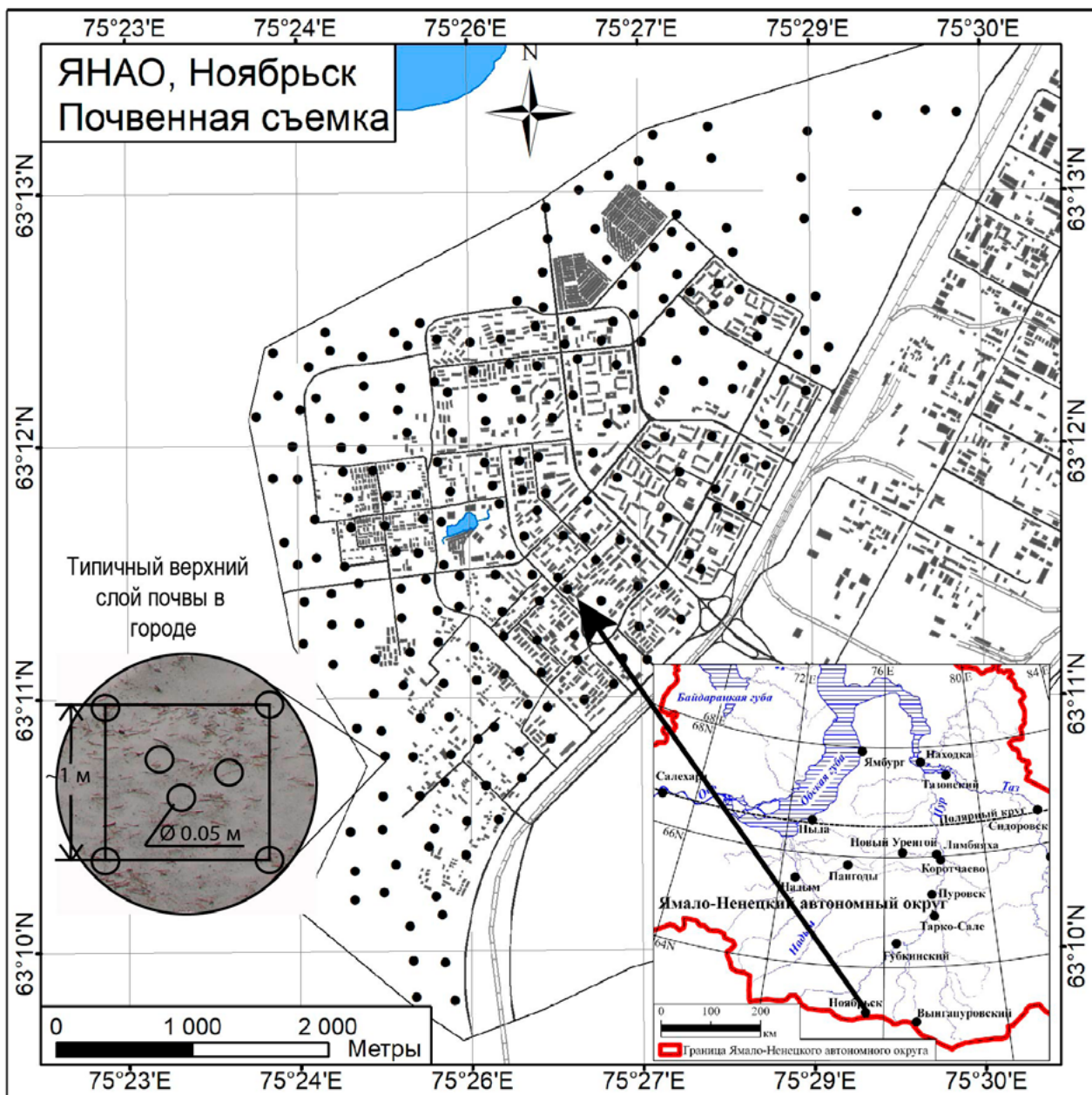


Рис. 1. Схема отбора проб

### 2.3. Подготовка данных для моделирования

Географические координаты были входными данными для моделирования, а содержание элемента – выходными. Для очистки исходных данных от выбросов использовался метод линейной интерполяции соседних значений [26]. Пространственные координаты и содержание Cr масштабировали от 0 до 1.

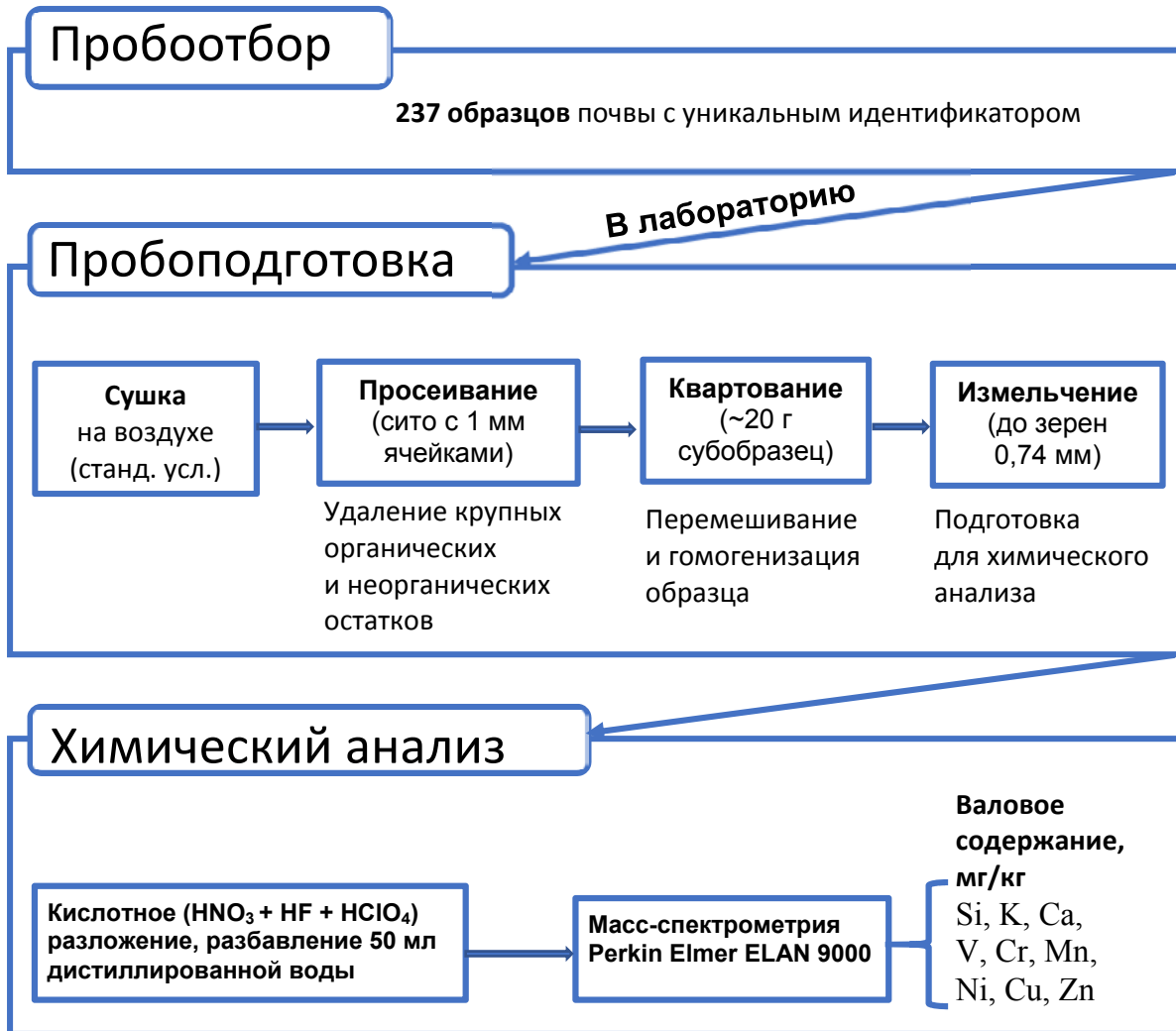


Рис. 2. Этапы пробоподготовки и химического анализа

### 2.4. МЛП

В качестве эталонной модели искусственной нейронной сети использовался многослойный персептрон. Построение модели МЛП заключалось в выборе количества нейронов внутри скрытого слоя. В процессе обучения МЛП связи между нейронами устанавливались путем присвоения весов, обновлений веса и значений смещения в соответствии с функцией потерь как наименьшей суммы квадратов ошибок в обучающей выборке (алгоритм обучения Левенберга – Марквардта). Мы использовали логистическую функцию активации.

Структура МЛП была выбрана с помощью компьютерного моделирования на основе минимального RMSE для всех данных (237 точек). Входной слой МЛП состоял из двух нейронов (пространственные координаты  $x$  и  $y$ ). МЛП имел один скрытый слой с числом нейронов от 1 до 10. Выходной слой МЛП включал один нейрон (элемент содержимого).

## 2.5. Оценка точности модели

Была использована среднеквадратическая ошибка RMSE (1) для проверки точности прогнозирования между прогнозом и набором необработанных данных.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (p(i) - o(i))^2}{n}}, \quad (1)$$

где  $p(i)$  – прогнозируемые данные,  
 $o(i)$  – измеренные данные,  
 $n$  – количество точек.

## 2.6. Выбор обучающего подмножества для построения МЛП

### 2.6.1. Случайный алгоритм

Полностью случайный (Random) алгоритм, при котором набор необработанных данных случайным образом разбивали на обучающее и тестовое подмножества в соотношении 70 %/30 % соответственно.

Каждую структуру сети (10 сетей) с соответствующим количеством нейронов в скрытом слое обучали десять раз и выбирали лучшую (с минимальным RMSE). Для каждого разделения получали 100 сетей: от 1 до 10 нейронов и десятикратное обучение. Среди всех этих сетей выбирали лучшую с минимальным RMSE.

### 2.6.2. Трехшаговый (3step) алгоритм разделения данных

Исходные данные многократно делили случайным образом на обучающее и тестовое подмножества в соотношении 70 %/30 % соответственно. Количество способов разделения равно числу сочетаний без повторения. Алгоритм разделения исходных данных состоит из трех шагов.

1. Набор исходных данных 1 000 раз случайным образом разделили на два непересекающихся множества, обучающую и тестовую выборки в соотношении 70 %/30 % соответственно. Таким образом, получили 1 000 обучающих и 1 000 тестовых подмножеств.

2. Для каждого случайного обучающего подмножества построили 1 000 сетей. Для каждой обученной сети определялась среднеквадратическая ошибка (RMSE) прогноза тестового набора.

3. Отобрали 100 (10 % от общего числа или 0,1-квантиль) сетей с наименьшим RMSE. Частоты попаданий в обучающую выборку вычислялись путем суммирования попаданий каждой точки пространства по выбранным сетям. Для включения в обучающее подмножество выбирались точки, частота попаданий которых в обучающее подмножество превышала 75 %.

### 2.6.3. Четырехшаговый (4step) алгоритм построения обучающего подмножества

Алгоритм построения обучающего подмножества 4step, подобно трехшаговому алгоритму, основан на делении по числу сочетаний без повторения, но отличается построением индивидуальной для каждой точки оценки репрезентативности по числу попаданий в обучающий набор.

1. Исходные данные 100 раз случайным образом разделили на обучающее и тестовое подмножества с соотношением 70 %/30 % соответственно. Обучающее подмножество включало 166 точек, тестовое 71 точку. Получили 100 разбиений на два непересекающихся подмножества.

2. Для каждого разделения построили десять моделей МЛП. Всего обучили 1 000 МЛП. Для каждого разделения выбирали один из десяти MLP с минимальной RMSE для оценки репрезентативности точки выборки. Таким образом, на 100 разделений было получено 100 МЛП, для которых рассчитывали среднеквадратичные ошибки обучающих и тестовых подмножеств.

3. Каждой точке выборки присвоили набор наилучших сетей, в которых она участвовала в обучении, и построили гистограммы распределения RMSE для обучающего, тестового и общего множеств.

4. Для каждой точки выборки рассчитали базовые статистики RMSE обучающего, тестового и общего подмножества для сетей, в которых точка участвовала в обучении. Репрезентативность точки оценивалась путем сравнения средних значений и стандартных отклонений.

#### **2.6.4. Алгоритм пространственного квотирования**

Алгоритм пространственного квотирования (Space&max&min quote) основан на идее, что каждый участок изучаемой территории может быть представлен квотой отобранных на ней проб. Алгоритм включал следующие шаги.

1. Район съемки оконтуривали выпуклым многоугольником так, что геодезическая линия, проведенная между любыми двумя точками опробования, находилась внутри этого многоугольника. Оконтуривание выполнялось путем соединения граничных точек. Такая процедура удовлетворяет условию интерполяции, поскольку любой прогноз находится внутри этого многоугольника.

2. Полученный многоугольник разбивали на участки, включающие одинаковое количество наблюдений – точек опробования. В обучающее подмножество обязательно включали граничные точки полигона, максимальные и минимальные значения из каждой пространственной квоты. Количество пространственных квот полагали не менее четырех. Это согласуется с выбранными географическими направлениями. При наличии ярко выраженных географических особенностей (например, резкое изменение растительности, наличие ярко выраженного рельефа и т. п.) эту зону выделяли в отдельный пространственный участок. Объем пространственной квоты ограничен снизу необходимым количеством точек тестового подмножества. Из общих статистических соображений желательно, чтобы это подмножество содержало не менее 30 точек для надежной оценки точности интерполятора. Если объем данных позволяет, то количество пространственных участков можно увеличить.

3. Из каждого участка случайным образом добирали для обучающего подмножества точки так, чтобы их доля составляла 70 %.

Окончательное обучающее подмножество состояло из граничных точек, максимальных и минимальных значений из каждого участка и случайных добавлений из каждого участка до 70 % от общего числа точек. Оставшиеся 30 % точек составляли тестовое подмножество.

### 2.6.5. Алгоритм пространственно-вероятностного квотирования

Алгоритм пространственно-вероятностного квотирования (Space&quartile quote) повторял пространственное квотирование исходных данных, но с учетом возможного изменения значений моделируемой переменной.

1 и 2 шага одинаковы из алгоритма Space&max&min.

3. После второго шага пространственного квотирования исходных данных оставшиеся значения моделируемой переменной из каждого участка были разбиты на вероятностные квоты (в обучающую выборку попали граничные точки, максимальное и минимальное значения). Желательно, чтобы количество вероятностных квот было не менее четырех для минимально необходимой репрезентативности функции плотности вероятности (левый, правый хвост и левый, правый центр).

4. 70 % значений из каждой квоты вероятностей были случайным образом выбраны для обучающего подмножества. Возможны варианты, в которых граничные точки и максимальное и минимальное значения могут совпадать. В этих случаях в обучающее подмножество попадает большее количество «случайных» значений. В тестовую выборку вошли оставшиеся 30 % точек. В общем случае количество вероятностных квот не может превышать количество тестовых случаев, т. е. каждая вероятностная квота должна быть представлена хотя бы одной точкой.

## 3. Результаты

Описательная статистика содержания хрома представлена в табл. 1. Гистограмма распределения содержаний хрома на выбранной территории приведена на рис. 3. Общее содержание Cr на городском фоне находится в пределах от 17 до 140 мг/кг, что не превышает кларк (по Уралу) [2, 27]. Известно, что общее содержание Cr в подзолах в почвах мира находится в пределах от 2,6 до 200 мг/кг [28]. Распределение хрома скошено вправо, с правым хвостом. Средние значения больше, чем (справа) медианы. Тест Шапиро – Уилка показал, что распределение отличается от нормального.

Таблица 1. Характеристика распределения хрома в верхнем слое почвы г. Ноябрьска

Содержание, мг/кг					Коэффициент вариации, %	Асимметрия	Эксцесс	p-значение Шапиро – Уилка
Минимум	Максимум	Среднее	CO <sup>*</sup> )	Медиана				
17	140	63	23	60	37	0,8	0,6	0,00

CO<sup>\*</sup>) – стандартное отклонение.



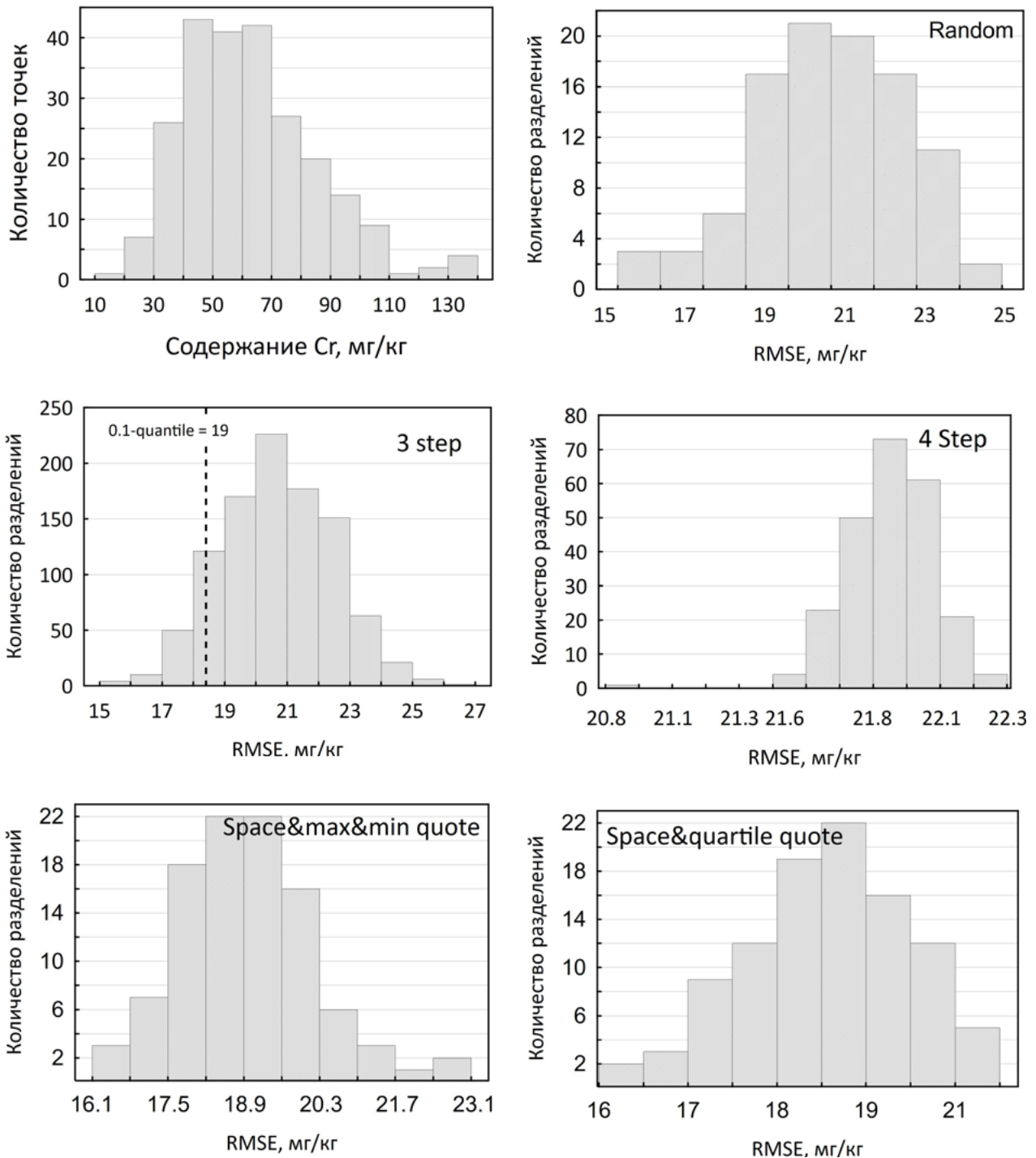


Рис. 3. Распределение исходных данных распределения признака и RMSE для разных алгоритмов разделения данных на обучающее и тестовое подмножества

В табл. 2 представлены статистические характеристики RMSE для разных способов разделения данных. На рис. 2 приведены гистограммы распределений RMSE для разных способов выбора обучающего подмножества. На рис. 3 приведено сравнение разных способов выбора обучающего множества по средним значениям точности модели RMSE. Наилучший результат по среднему значению RMSE показывает алгоритм пространственно-вероятностного квотирования. Интерес представляет четырехшаговый алгоритм построения обучающего подмножества, при самой высокой средней ошибке из представленных, распределение RMSE отличается от нормального и имеет самые низкие стандартное отклонение и коэффициент вариации.



Таблица 2. Характеристика распределения RMSE для разных алгоритмов

Алгоритм	RMSE, мг/кг							Коэффициент вариации, %	Асимметрия	Экцесс
	Среднее	Медиана	Минимум	Максимум	CO	Среднее – CO	Среднее + CO			
Random	20,6	20,6	15,3	24,9	1,9	18,7	22,5	9,2	– 0,36	– 0,05
3step	20,6	20,6	15,4	26,5	1,8	18,8	22,4	8,5	0,04	– 0,21
4step	<b>21,9</b>	<b>21,9</b>	<b>20,8</b>	<b>22,3</b>	<b>0,2</b>	<b>21,7</b>	<b>22,1</b>	<b>0,8</b>	<b>– 1,01</b>	<b>6,26</b>
Space&max&min quote	18,9	18,9	16,1	23,0	1,3	17,6	20,2	6,7	0,59	0,95
Space&quartile quote	<b>18,8</b>	<b>18,9</b>	<b>15,8</b>	<b>21,1</b>	<b>1,1</b>	<b>17,7</b>	<b>19,9</b>	<b>5,8</b>	<b>– 0,34</b>	<b>– 0,19</b>

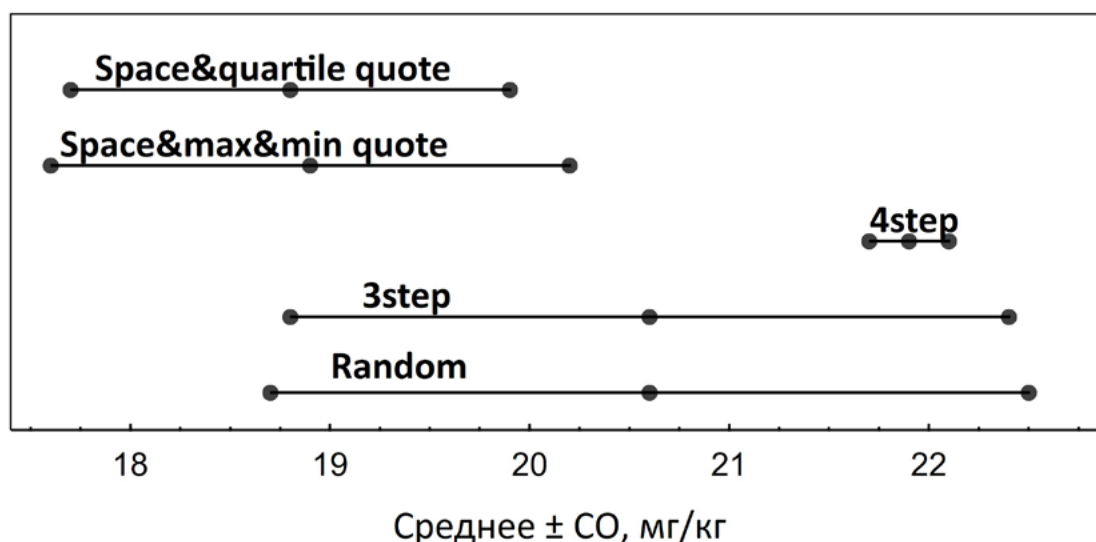


Рис. 4. Вариативность среднего для разных алгоритмов

Алгоритмы Space&quartile quote и Space&max&min quote относятся к детерминированным с элементом случайности, поскольку часть точек обучающего подмножества зафиксирована. 3step и 4step алгоритмы основаны на разделении исходных данных только случайным образом.

На рис. 5 представлена схема расположения точек, включение которых в обучающее подмножество обеспечивает для разных алгоритмов наименьшую RMSE. На рис. 5 представлены три списка точек для обучения: для детерминированного (Space quote), 3step и 4step алгоритмов. Для детерминированных Space&quartile

quote и Space&max&min quote способов обязательные для включения в обучающее множество точки совпадают по условиям построения алгоритмов. Наилучшие для обучения точки для 3step и 4step алгоритмов отличаются. Точки обучающего подмножества для разных алгоритмов не являются общими.

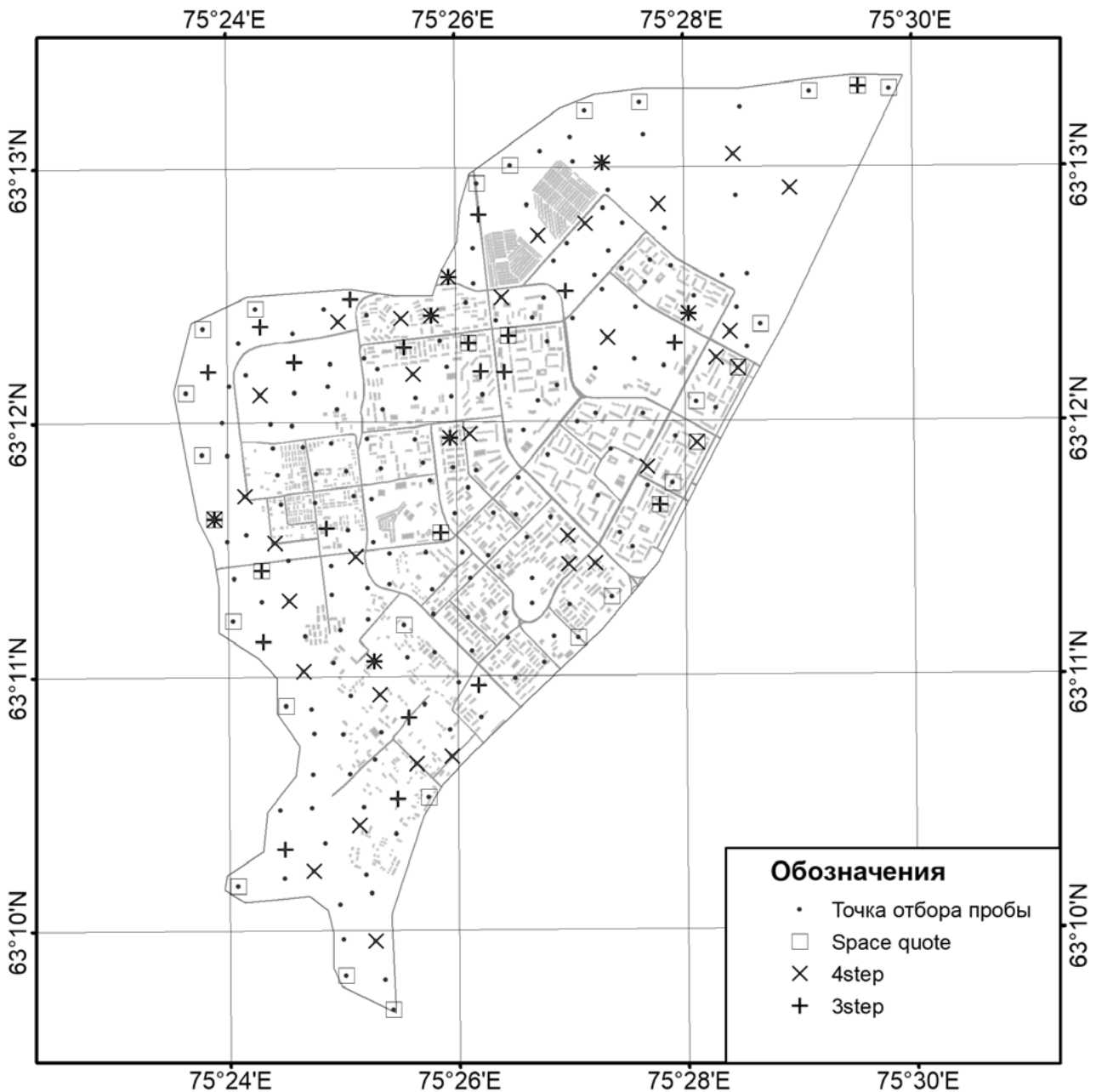


Рис. 5. Расположение репрезентативных точек обучающего подмножества для разных алгоритмов разделения данных

#### 4. Обсуждение

Модель МЛП, использующая контролируемые типы разбиения, оказалась более точной, чем МЛП, обученная на случайном разбиении. Алгоритм Space&quartile quote показал наиболее точные результаты. Относительно небольшое количество разных точек в этих обучающих подмножествах может объяснить небольшую разницу между двумя детерминированными типами разбиения. Модели Space&quartile и Space&max&min с контролируемым или детерминированным разделением более

точные. При этом распределение ошибок RMSE этих моделей уже, чем в моделях, основанных на случайном разбиении. МЛП с пространственно-вероятностным квотированием исходных данных с учетом разброса значений Space&quartile оказался несколько более точным, чем МЛП только с пространственным квотированием Space&max&min. Разница между алгоритмами Space&quartile и Space&max&min была незначительной (менее 5 %). Для относительно небольших выборок процедура разделения на обучающее и тестовое подмножества может быть практически полностью определена. Это упрощает расчет модели при неуменьшении точности модели. Для таких детерминированных способов разделения одна реализация будет более точной (или сравнимой по точности), чем лучшая из сотен сетей, обучавшихся на случайно разделенной выборке. Алгоритмы, построенные на многократном случайном разделении, обеспечивают более высокую среднюю RMSE при меньшем отклонении.

## 5. Выводы

1. Точность модели ИНС для предсказания пространственного распределения признака в экологических задачах зависит от способа разделения исходных данных на обучающее и тестовое подмножества.
2. Детерминированные алгоритмы, основанные на квотировании исходных данных при разделении на обучающее и тестовое подмножества, обеспечивают меньшую среднюю ошибку модели, но разброс этой ошибки будет составлять 5–7 %.

## 6. Список литературы

1. *Demyanov, V. A special issue on data science for geosciences / V. Demyanov, E. Gloaguen, M. Kanevski // Math. Geosci. – 2020. – Vol. 52. – P. 1–3. <https://doi.org/10.1007/s11004-019-09846-0>*
2. *Геохимия окружающей среды / Ю. Е. Саэт, Б. А. Ревич, Е. П. Янин [и др.]. – М. : Недра, 1990. – 335 с. – ISBN 5-247-01127-9. – EDN XDXBQN.*
3. *A geochemical survey of heavy metals in agricultural and background soils of the Isfahan industrial zone Iran / A. Esmaeili, F. Moore, B. Keshavarzi [et al.] // Catena. – 2017. – Vol. 121. – P. 88–98.*
4. *Spatial distribution and pollution assessment of heavy metals in urban soils from southwest China / G. H. Guo, F. Wu, F. Xie [et al.] // J. Env. Sci. – 2012. – Vol. 24(3). – P. 410–418.*
5. *High variation topsoil pollution forecasting in the Russian Subarctic: Using artificial neural networks combined with residual kriging / D. A. Tarasov, A. G. Buevich, A. P. Sergeev [et al.] // Applied Geochemistry. – 2018. – Vol. 88. – P. 188–197. – DOI 10.1016/j.apgeochem.2017.07.007. – EDN XXSNEL.*
6. *Zhong, L. Soil properties: Their prediction and feature extraction from the LUCAS spectral library using deep convolutional neural networks / X. Guo, Zh. Xu, M. Ding // Geoderma. – 2021. – Vol. 402. – P. 115366. <https://doi.org/10.1016/j.geoderma.2021.115366>*
7. *Nonlinear variable selection for artificial neural networks using particle mutual information / R. J. May, H. R. Maier, G. D. Dandy [et al.] // Environmental Modelling & Software. – 2008. – Vol. 23 (10–11). – P. 1312–1326.*
8. *Shaker, R. R. Exploring non-linear relationships between landscape and aquatic ecological condition in southern Wisconsin: A GWR and ANN approach / R. R. Shaker, T. J. Ehlinger // International J. of Applied Geospatial Research. – 2014. – Vol. 5(4). – P. 1–20.*

9. *Kanevski, M.* Mapping of soil contamination by using artificial neural networks and multivariate geostatistics / M. Kanevski, V. Demyanov, M. Maignan // *Lecture Notes in Computer Science*. – 1997. – Vol. 1327. – P. 1125–1130. – DOI 10.1007/bfb0020304. – EDN WQHJZF.
10. The Effect of Splitting of Raw Data into Training and Test Subsets on the Accuracy of Predicting Spatial Distribution by a Multilayer Perceptron / E. M. Baglaeva, A. P. Sergeev, A. V. Shichkin [et al.] // *Mathematical Geosciences*. – 2020. – Vol. 52. – No 1. – P. 111–121. – DOI 10.1007/s11004-019-09813-9. – EDN HBCZYX.
11. Почвы и ноосфера: Монография. Научное электронное издание / Е. Р. Абрамова, И. А. Аднан, А. В. Базаров [и др.]; Министерство науки и высшего образования РФ, Дальневосточный федеральный университет, Российская академия наук, Дальневосточное отделение, ФНЦ биоразнообразия наземной биоты Восточной Азии, Общество почвоведов им. В. В. Докучаева, Дальневосточное отделение общества почвоведов им. В. В. Докучаева, Far Eastern Climate Smart Lab. – Владивосток: Дальневосточный федеральный университет, 2019. – 409 с. – ISBN 978-5-7444-4707-6. – EDN LMCWKD.
12. A back propagation neural network model optimized by mind evolutionary algorithm for estimating Cd, Cr, and Pb concentrations in soils using Vis-NIR diffuse reflectance spectroscopy / X. Wang, Sh. An, Y. Xu [et al.] // *Appl. Sci.* – 2020. – Vol. 10(51). – P. 1–17. <https://doi.org/10.3390/app10010051>.
13. Coordinate Transformation between Global and Local Data Based on Artificial Neural Network with K-Fold Cross-Validation in Ghana / Y. Y. Ziggah, H. Youjian, A. R. Tierra [et al.] // *Earth Sciences Research J.* – 2019. – Vol. 23(1). – PP. 67–77. <https://doi.org/10.15446/esrj.v23n1.63860>.
14. *Nath, A.* The role of pertinently diversified and balanced training as well as testing data sets in achieving the true performance of classifiers in predicting the antifreeze proteins / A. Nath, K. Subbiah // *Neurocomputing*. – 2018. – Vol. 272. – P. 294–305. <http://dx.doi.org/10.1016/j.neucom.2017.07.004>.
15. *Челахов, В. М.* Метод кортежирования обучающей выборки для повышения точности нейросетевой аппроксимации данных / В. М. Челахов, К. В. Деркачев // *Современные проблемы науки и образования*. – 2012. – № 6. – С. 107. – EDN TODLRF.
16. *Бурый, Я. А.* Экстраполирующее обучение нейронных сетей / Я. А. Бурый, Д. И. Самаль // *Информатика*. – 2019. – Т. 16, № 1. – С. 86–92. – EDN PQWKJT.
17. *Донской, В. И.* Извлечение оптимизационных моделей из данных: применение нейронных сетей / В. И. Донской // *Таврический вестник информатики и математики*. – 2018. – № 2(39). – С. 71–89. – EDN YPPSLR.
18. *Пастухов, А. А.* Применение самоорганизующихся карт Кохонена для формирования представительской выборки при обучении многослойного персептрона / А. А. Пастухов, А. А. Прокофьев // *Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Физико-математические науки*. – 2016. – № 2(242). – С. 95–107. – DOI 10.5862/JPM.242.11. – EDN WAEDIH.
19. *Malof, J. M.* How do we choose the best model? The impact of cross-validation design on model evaluation for buried threat detection in ground penetrating radar / J. M. Malof, D. Reichman, L. M. Collins // *Proceedings Volume 10628, Detection and Sensing of Mines, Explosive Objects, and Obscured Targets*. – 2018. – Vol. XXIII. – 106280C. <https://doi.org/10.1117/12.2305793>.
20. *Kramm, T.* Assessing the influence of environmental factors and datasets on soil type prediction with two machine learning algorithms in a heterogeneous area in the Rur catch-

- ment, Germany / T. Kramm, D. Hoffmeister // *Geoderma Regional*. – 2020. – Vol. 22. – e00316. <https://doi.org/10.1016/j.geodrs.2020.e00316>.
21. *Fernandez Jaramillo, J. M.* Sample selection via angular distance in the space of the arguments of an artificial neural network / J. M. Fernandez Jaramillo, R. Mayerle // *Computers and Geosciences*. – 2018. – Vol. 114. – P. 98–106. <https://doi.org/10.1016/j.cageo.2018.02.003>.
  22. *Sakizadeh, M.* Support vector machine and artificial neural network to model soil pollution: a case study in Semnan Province, Iran / M. Sakizadeh, R. Mirzaei, H. Ghorbani // *Neural. Comput. & Applic.* – 2017. – Vol. 28. – P. 3229–3238. <https://ezproxy.urfu.ru:3055/10.1007/s00521-016-2231-x>.
  23. Forecasting the spatiotemporal variability of soil CO<sub>2</sub> emissions in sugarcane areas in southeastern Brazil using artificial neural networks / L. P. S. Freitas, M. L. M. Lopes, L. B. Carvalho [et al.] // *Environ. Monit. Assess.* – 2018. – Vol. 190. – P. 741. <https://doi.org/10.1007/s10661-018-7118-0>
  24. Combining spatial autocorrelation with machine learning increases prediction accuracy of soil heavy metals / A. P. Sergeev, A. G. Buevich, E. M. Baglaeva [et al.] // *Catena*. – 2019. – Vol. 174. – P. 425–435. – DOI 10.1016/j.catena.2018.11.037. – EDN AZBEVO.
  25. *Добровольский, Г. В.* География почв / Г. В. Добровольский, И. С. Урусевская. – М. : Изд-во Московского университета, 1984. – 416 с. – EDN TSHOUB.
  26. Clean Outlier Data of Matlab. [www.mathworks.com/help/matlab/ref/cleanoutlierdata.html](http://www.mathworks.com/help/matlab/ref/cleanoutlierdata.html)
  27. Краткий справочник по геохимии / Г. В. Войткевич, А. Е. Мирошников, А. С. Поваренных [и др.]. – М. : Недра, 1977. – 180 с.
  28. *Kabata-Pendias, A.* Trace elements in soils and plants / A. Kabata-Pendias. – Taylor and Francis Group CRC Press, 2011. <https://doi.org/10.1201/b10158>.

## Сведения об авторах:

**Баглаева Елена Михайловна**, канд. ф.-м. н., старший научный сотрудник Института промышленной экологии УрО РАН, г. Екатеринбург, Россия. Эл. почта: [elenbaglaeva@gmail.com](mailto:elenbaglaeva@gmail.com).

**Сергеев Александр Петрович**, канд. ф.-м. н., заведующий Лаборатории физики и экологии Института промышленной экологии УрО РАН, г. Екатеринбург, Россия.

**Шичкин Андрей Васильевич**, научный сотрудник Института промышленной экологии УрО РАН, г. Екатеринбург, Россия

**Бувич Александр Геннадьевич**, научный сотрудник Института промышленной экологии УрО РАН, г. Екатеринбург, Россия

**Субботина Ирина Евгеньевна**, канд. ф.-м. н., научный сотрудник Института промышленной экологии УрО РАН, г. Екатеринбург, Россия

**Буторова Анастасия Сергеевна**, инженер-исследователь Института промышленной экологии УрО РАН, г. Екатеринбург, Россия

# COMPARATIVE ANALYSIS OF DATA SPLITTING ALGORITHMS FOR TRAINING ARTIFICIAL NEURAL NETWORKS IN MODELING THE SPATIAL DISTRIBUTION OF HEAVY METALS IN THE TOPSOIL

E. M. Baglaeva, A. P. Sergeev, A. V. Shichkin, A. G. Buevich, I. E. Subbotina,  
A. S. Butorova

*Institute of Industrial Ecology, Ural Branch of the Russian Academy of Sciences,  
Ekaterinburg, Russia*

*A comparison of five algorithms for selecting a training subset for artificial neural networks to interpolate the spatial distribution of a feature according to environmental components screenings is presented. The initial data were the contents of chromium in the topsoil in the residential areas of the city of Noyabrsk (YNAO, Russia). The spatial distributions of the element content in the topsoil were interpolated by a multilayer perceptron (MLP). Root mean squared error (RMSE) statistics were calculated as accuracy metrics for each algorithm. A smaller error is shown by the Space&quartile quote algorithm based on the spatial-probabilistic quoting of the initial data. Taking into account the information about the error statistics of the algorithms for splitting the original set, it is possible to improve the accuracy of the artificial neural network model in solving environmental problems.*

**Key words:** MLP; splitting data; spatial distribution; environment modelling; representativeness.

## References

1. *Demyanov, V. A special issue on data science for geosciences / V. Demyanov, E. Gloaguen, M. Kanevski // Math. Geosci. – 2020. – Vol. 52. – P. 1–3. <https://doi.org/10.1007/s11004-019-09846-0>*
2. *Geochemistry of the environment / Yu. E. Saet, B. A. Revich, E. P. Yanin [et al.]. – Moscow : Nedra Publishing House, 1990. – 335 p. – ISBN 5-247-01127-9. – EDN XDXBQN.*
3. *Esmaeili, A. A geochemical survey of heavy metals in agricultural and background soils of the Isfahan industrial zone Iran / A. Esmaeili, F. Moore, B. Keshavarzi [et al.] // Catena. – 2017. – Vol. 121. – P. 88–98.*
4. *Spatial distribution and pollution assessment of heavy metals in urban soils from southwest China / G. H. Guo, F. Wu, F. Xie [et al.] // J. Env. Sci. – 2012. – Vol. 24(3). – P. 410–418.*
5. *High variation topsoil pollution forecasting in the Russian Subarctic: Using artificial neural networks combined with residual kriging / D. A. Tarasov, A. G. Buevich, A. P. Sergeev [et al.] // Applied Geochemistry. – 2018. – Vol. 88. – P. 188–197. – DOI 10.1016/j.apgeochem.2017.07.007. – EDN XXSNEL.*

6. *Zhong, L.* Soil properties: Their prediction and feature extraction from the LUCAS spectral library using deep convolutional neural networks / X. Guo, Zh. Xu, M. Ding // *Geoderma*. – 2021. – Vol. 402. – P. 115366. <https://doi.org/10.1016/j.geoderma.2021.115366>
7. Nonlinear variable selection for artificial neural networks using particle mutual information / R. J. May, H. R. Maier, G. D. Dandy [et al.] // *Environmental Modelling & Software*. – 2008. – Vol. 23 (10–11). – P. 1312–1326.
8. *Shaker, R. R.* Exploring non-linear relationships between landscape and aquatic ecological condition in southern Wisconsin: A GWR and ANN approach / R. R. Shaker, T. J. Ehlinger // *International J. of Applied Geospatial Research*. – 2014. – Vol. 5(4). – P. 1–20.
9. *Kanevski, M.* Mapping of soil contamination by using artificial neural networks and multivariate geostatistics / M. Kanevski, V. Demyanov, M. Maignan // *Lecture Notes in Computer Science*. – 1997. – Vol. 1327. – P. 1125–1130. – DOI 10.1007/bfb0020304. – EDN WQHJZF.
10. The Effect of Splitting of Raw Data into Training and Test Subsets on the Accuracy of Predicting Spatial Distribution by a Multilayer Perceptron / E. M. Baglaeva, A. P. Sergeev, A. V. Shichkin [et al.] // *Mathematical Geosciences*. – 2020. – Vol. 52. – No 1. – P. 111–121. – DOI 10.1007/s11004-019-09813-9. – EDN HBCZYX.
11. Soils and noosphere: Monograph. Scientific electronic edition / E. R. Abramova, I. A. Adnan, A. V. Bazarov [et. al.]; Ministry of Science and Higher Education of the Russian Federation, Far Eastern Federal University, Russian Academy of Sciences, Far Eastern Branch, Federal Scientific Center for Biodiversity of Terrestrial Biota of East Asia, Society of Soil Scientists named after V. V. Dokuchaev, Far Eastern Branch of the Society of Soil Scientists named after A. I. V. V. Dokuchaeva, Far Eastern Climate Smart Lab. – Vladivostok: Far Eastern Federal University, 2019. – 409 p. – ISBN 978-5-7444-4707-6. – EDN LMCWKD.
12. A back propagation neural network model optimized by mind evolutionary algorithm for estimating Cd, Cr, and Pb concentrations in soils using Vis-NIR diffuse reflectance spectroscopy / X. Wang, Sh. An, Y. Xu [et al.] // *Appl. Sci*. – 2020. – Vol. 10(51). – P. 1–17. <https://doi.org/10.3390/app10010051>.
13. Coordinate Transformation between Global and Local Data Based on Artificial Neural Network with K-Fold Cross-Validation in Ghana / Y. Y. Ziggah, H. Youjian, A. R. Tierra [et al.] // *Earth Sciences Research J*. – 2019. – Vol. 23(1). – PP. 67–77. <https://doi.org/10.15446/esrj.v23n1.63860>.
14. *Nath, A.* The role of pertinently diversified and balanced training as well as testing data sets in achieving the true performance of classifiers in predicting the antifreeze proteins / A. Nath, K. Subbiah // *Neurocomputing*. – 2018. – Vol. 272. – P. 294–305. <http://dx.doi.org/10.1016/j.neucom.2017.07.004>.
15. *Chelakhov, V. M.* Method of training sample tuple to improve the accuracy of neural network approximation of data / V. M. Chelakhov, K. V. Derkachev // *Modern problems of science and education*. – 2012. – No. 6. – P. 107. – EDN TODLRF.
16. *Bury, Ya. A.* Extrapolating training of neural networks / Ya. A. Bury, D. I. Samal // *Informatics*. – 2019. – T. 16, No. 1. – S. 86–92. – EDN PQWKJT.
17. *Donskoy, V. I.* Extraction of optimization models from data: the use of neural networks / V. I. Donskoy // *Tauride Bulletin of Informatics and Mathematics*. – 2018. – No. 2 (39). – S. 71–89. – EDN YPPSLR.
18. *Pastukhov, A. A.* The use of self-organizing Kohonen maps for the formation of a representative sample when training a multilayer perceptron / A. A. Pastukhov, A. A. Prokofiev // *Scientific and technical statements of the St. Petersburg State Polytechnic University*.

- Physical and mathematical sciences. – 2016, No. 2 (242). – P. 95–107. – DOI 10.5862/JPM.242.11. – EDN WAEDIH.
19. *Malof, J. M.* How do we choose the best model? The impact of cross-validation design on model evaluation for buried threat detection in ground penetrating radar / J. M. Malof, D. Reichman, L. M. Collins // Proceedings Volume 10628, Detection and Sensing of Mines, Explosive Objects, and Obscured Targets. – 2018. – Vol. XXIII. – 106280C. <https://doi.org/10.1117/12.2305793>.
  20. *Kramm, T.* Assessing the influence of environmental factors and datasets on soil type prediction with two machine learning algorithms in a heterogeneous area in the Rur catchment, Germany / T. Kramm, D. Hoffmeister // Geoderma Regional. – 2020. – Vol. 22. – e00316. <https://doi.org/10.1016/j.geodrs.2020.e00316>.
  21. *Fernandez Jaramillo, J. M.* Sample selection via angular distance in the space of the arguments of an artificial neural network / J. M. Fernandez Jaramillo, R. Mayerle // Computers and Geosciences. – 2018. – Vol. 114. – P. 98–106. <https://doi.org/10.1016/j.cageo.2018.02.003>.
  22. *Sakizadeh, M.* Support vector machine and artificial neural network to model soil pollution: a case study in Semnan Province, Iran / M. Sakizadeh, R. Mirzaei, H. Ghorbani // Neural. Comput. & Applic. – 2017. – Vol. 28. – P. 3229–3238. <https://ezproxy.urfu.ru:3055/10.1007/s00521-016-2231-x>.
  23. Forecasting the spatiotemporal variability of soil CO<sub>2</sub> emissions in sugarcane areas in southeastern Brazil using artificial neural networks / L. P. S. Freitas, M. L. M. Lopes, L. B. Carvalho [et al.] // Environ. Monit. Assess. – 2018. – Vol. 190. – P. 741. <https://doi.org/10.1007/s10661-018-7118-0>
  24. Combining spatial autocorrelation with machine learning increases prediction accuracy of soil heavy metals / A. P. Sergeev, A. G. Buevich, E. M. Baglaeva [et al.] // Catena. – 2019. – Vol. 174. – P. 425–435. – DOI 10.1016/j.catena.2018.11.037. – EDN AZBEVO.
  25. *Dobrovolsky, G. V.* Geography of soils / G. V. Dobrovolsky, I. S. Urusevskaya. – M. : Publishing House of the Moscow State University, 1984. – 416 p. – EDN TSHOUB.
  26. Clean Outlier Data of Matlab. [www.mathworks.com/help/matlab/ref/cleanoutlierdata.html](http://www.mathworks.com/help/matlab/ref/cleanoutlierdata.html)
  27. Brief reference book on geochemistry / G. V. Voitkevich, A. E. Miroshnikov, A. S. Cookery [et al.]. – M. : Nedra, 1977. – 180 p.
  28. *Kabata-Pendias, A.* Trace elements in soils and plants / A. Kabata-Pendias. – Taylor and Francis Group CRC Press, 2011. <https://doi.org/10.1201/b10158>.



# СОВМЕСТНОЕ ДЕЙСТВИЕ БИНАРНЫХ ФАКТОРОВ И РАВЕНСТВО РИСКОВ

Ю. В. Нагребецкая, А. И. Фаткуллина

Уральский Федеральный университет им. первого Президента России  
Б. Н. Ельцина, г. Екатеринбург, Россия

*В работе исследуются широко применяемые в медико-биологических науках модели взаимодействия двух и трех независимых бинарных факторов в случае бинарного отклика. Основное внимание уделено формализации понятия «взаимодействие» и его связи с условием равенства разностей рисков. Показано, что это условие можно получить из приравнивания нулю всех членов взаимодействия в их представлении, применяемом в дисперсионном анализе. Получена и решена система равенства рисков для двух и трех факторов, а также показана ее инвариантность относительно действия группы симметрий эксперимента. Приведена интерпретация решения системы равенства рисков: в случае двух факторов эффект от их воздействия равен сумме эффектов от изолированного воздействия каждого фактора в отдельности; в случае трех факторов (а) эффект от воздействия всех этих факторов равен сумме эффектов от изолированного воздействия каждого фактора в отдельности; (б) эффект от воздействия любых двух факторов равен сумме эффектов от изолированного воздействия каждого фактора в отдельности при отсутствии воздействия третьего. Кроме того, проведено исследование решения равенства рисков для двух и трех факторов в рамках детерминистической модели, приводящей к булевой формализации этой модели. Доказано, что для значений булевой функции, представляющей данный отклик, система равенства рисков и для двух, и для трех факторов выполняется тогда и только тогда, когда эта булева функция существенно зависит не более чем от одной переменной.*

*Показано, что в рамках детерминистической модели понятие «совместное действие» бинарных факторов, введенное и исследованное в работах первого автора, полнее описывает взаимодействие этих факторов, чем нарушение условия равенства разностей рисков.*

**Ключевые слова:** условие равенства разностей рисков; детерминистическая модель; бинарная теория достаточных причин; отклик; совместное действие бинарных факторов; булева функция; группа автоморфизмов; действие группы.

## 1. Введение

Проблема описания взаимодействия составляет одну из важнейших задач медико-биологических наук. Помимо теоретического интереса, эта задача имеет и очевидный экономический аспект. Например, если известно, что изолированное присутствие некоторых агентов приводит к увеличению частоты некоторого заболевания (для простоты будем считать, что одного и того же заболевания для всех аген-

тов), то может оказаться, что совместное воздействие этих агентов имеет другую распространенность этого заболевания, чем та, которую исследователь ожидает на основании информации об изолированных эффектах. Таким образом, в зависимости от того, больше или меньше фактическая заболеваемость, чем ожидаемая, ресурсы, необходимые для купирования ситуации многофакторного воздействия, будут больше или меньше прогнозируемых на основании однофакторных эффектов.

Эта проблема имеет значительную специфику при рассмотрении факторов, имеющих небольшое число уровней, например, один-два уровня воздействия. Такая дискретная ситуация не дает возможности применить непрерывные аппроксимации даже в качестве приближения. Поэтому необходимо рассматривать действующие факторы так, как они есть (без аппроксимаций) и вводить характеристики их воздействия на основе этих точных данных.

Один из первых вариантов численного описания особенностей совместного действия двухуровневых факторов и такого же отклика был предложен в работе [1] на основе представления причинного механизма воздействия в рамках так называемой теории достаточных причин (в работе [1] для нее предложено название *sufficient-component cause model*). Апеллируя непосредственно к эпидемиологической ситуации, автор вводит коэффициент ICDR (*Interaction Contrast of Disease Rates*), который можно понимать как комбинацию разностей рисков действующих факторов при различных сочетаниях их уровней. Здесь важно отметить, что вид этой комбинации таков, что позволяет трактовать смысл коэффициента ICDR непосредственно. Иначе говоря, арифметическая структура этого коэффициента определяется его интерпретацией как коэффициента взаимодействия. Это делает затруднительным распространение коэффициента ICDR за пределы двухфакторного двухуровневого случая.

Ниже предложена формальная процедура получения коэффициента Купмана для двух двухуровневых факторов и двухуровневого отклика, основанная на идеях дисперсионного анализа. Она может быть легко распространена на большее число двухуровневых факторов, при этом, однако, вместо одного уравнения, выражающего отсутствие взаимодействия, таких уравнений будет несколько, так что они образуют систему уравнений.

## 2. Система равенства рисков для двух факторов

Основное наблюдение, которое позволяет рассматривать случай двухуровневых факторов как представляющий самостоятельный интерес, состоит в следующем. Рассмотрим два фактора (две независимые переменные)  $A$ ,  $B$ , принимающих конечное число целых значений (уровней), а также результирующую (зависимую) переменную  $Y$ , вероятностная природа которой несущественна. Для каждой ячейки  $(i, j)$  обозначим теоретическое среднее значение отклика  $Y$  через  $\eta_{ij}$ . Имеющиеся данные позволяют провести анализ данных методами (двухфакторного) дисперсионного анализа [2–4] следующим образом. Пусть среднее значение отклика в ячейке  $(i, j)$  представлено в виде

$$\eta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij},$$

где  $\mu$  – среднее значение отклика во всей выборке,  
 $\alpha_i$  – главный эффект  $i$ -го уровня фактора  $A$ ,  
 $\beta_j$  – главный эффект  $j$ -го уровня фактора  $B$ ,  
 $\gamma_{ij}$  – член взаимодействия  $i$ -го уровня фактора  $A$  и  $j$ -го уровня фактора  $B$ .

Главные эффекты  $\alpha_i$  и  $\beta_j$ , член взаимодействия  $\gamma_{ij}$  и среднее значение отклика  $\mu$  можно представить в виде [4]:

$$\begin{aligned}\alpha_i &= \eta_{i\cdot} - \mu, \quad \beta_j = \eta_{\cdot j} - \mu, \\ \gamma_{ij} &= \eta_{ij} - \alpha_i - \beta_j - \mu, \quad \mu = \eta_{\cdot\cdot},\end{aligned}$$

где точка означает усреднение значений  $\eta_{ij}$  по соответствующему индексу. Тогда

$$\gamma_{ij} = \eta_{ij} - \eta_{i\cdot} - \eta_{\cdot j} + \eta_{\cdot\cdot} \quad (1)$$

Далее мы можем рассмотреть полностью бинарную ситуацию, в которой факторы и отклик являются двухуровневыми переменными, уровни которых будут кодироваться числами 0 и 1. При этом приведенные выше соотношения между введенными выше величинами будут сохраняться. Действительно, нигде в приведенных выше определениях не требовалось, чтобы результирующая переменная имела какую-то определенную природу. Таким образом, если действующие факторы и отклик являются бинарными переменными, то взаимодействие между уровнями  $A = i$ ,  $B = j$ ,  $i, j \in \{0, 1\}$  данных факторов в отклике также представляется равенством (1).

Важная особенность двухуровневой случайной переменной  $X$ , принимающей значение 0, 1 (бернуллиева случайная величина), состоит в том, что для нее выполняется равенство  $E[X] = P(X = 1) = p$ , где  $E[\dots]$  – оператор математического ожидания.

В рассматриваемой полностью бинарной ситуации введем обозначения

$$p_{ij} = P(Y = 1 \mid A = i, B = j), \quad i, j \in \mathbb{B},$$

где  $\mathbb{B} = \{0, 1\}$ .

Как обычно, в теоретическом анализе рассматриваются не выборочные характеристики, а их «истинные» значения, т. е. вместо усредненных величин берутся их математические ожидания, а вместо наблюдаемых частот появления события – вероятность их появления. Таким образом, условие отсутствия взаимодействия в данном случае имеет вид

$$\gamma_{ij} = p_{ij} - p_{i\cdot} - p_{\cdot j} + p_{\cdot\cdot} = 0, \quad i, j \in \mathbb{B}$$

Это равенство хорошо известно в эпидемиологии [5, 6] как условие равенства разностей рисков, или условие отсутствия эффекта от воздействия двух данных факторов. Трактовка этого равенства основана на представлении его в виде

$$p_{ij} - p_{\cdot\cdot} = (p_{i\cdot} - p_{\cdot\cdot}) - (p_{\cdot j} - p_{\cdot\cdot})$$

с последующим толкованием каждой разности  $p_{ij} - p_{\cdot\cdot}$ ,  $p_{i\cdot} - p_{\cdot\cdot}$  и  $p_{\cdot j} - p_{\cdot\cdot}$ .

Решим систему равенства рисков для двух факторов, принимая следующие обозначения:  $p_{i\cdot} = \frac{1}{2}(p_{i0} + p_{i1})$ ,  $p_{\cdot j} = \frac{1}{2}(p_{0j} + p_{1j})$ ,  $p_{\cdot\cdot} = \frac{1}{4}(p_{00} + p_{01} + p_{10} + p_{11})$

**Теорема 1.** Система из четырех уравнений (система равенства рисков)

$$p_{ij} - p_{i\cdot} - p_{\cdot j} + p_{\cdot\cdot} = 0, \quad i, j \in \mathbb{B} \quad (2)$$

равносильна единственному уравнению

$$p_{00} - p_{01} - p_{10} + p_{11} = 0 \quad (3)$$

**Доказательство** проводится непосредственными вычислениями.

Равенство (3) и представляет собой выражение коэффициента ICDR Купмана из работы [1].

**Определение 1.** Пусть значение уровня фактора  $A$  (фактора  $B$ ) равно 1 при воздействии фактора  $A$  (фактора  $B$ ) и равно 0 при отсутствии воздействия этого фактора. И пусть индексы  $i, j \in \mathbb{B}$  не равны нулю одновременно. Величину  $\Delta_{ij} = p_{ij} - p_{00}$  назовем *теоретическим эффектом* от воздействия фактора  $A$  (при отсутствии воздействия фактора  $A$ ), если  $i = 1$  ( $i = 0$ ), и воздействия фактора  $B$  (при отсутствии воздействия фактора  $B$ ), если  $j = 1$  ( $j = 0$ ). Разумеется, удобно предполагать, что  $\Delta_{00} = 0$ .

**Пример 1.**  $\Delta_{10}$  – теоретический эффект от воздействия фактора  $A$  при отсутствии воздействия фактора  $B$ ,  $\Delta_{01}$  – теоретический эффект от воздействия фактора  $B$  при отсутствии воздействия фактора  $A$ ,  $\Delta_{11}$  – теоретический эффект от воздействия факторов  $A$  и  $B$  одновременно.

Везде ниже, говоря об эффектах воздействия или отсутствии воздействия факторов, мы будем иметь в виду теоретические эффекты.

**Следствие 1.** Система (2) равенства рисков для двух факторов равносильна уравнению

$$\Delta_{11} = \Delta_{01} + \Delta_{10} \quad (4)$$

Равенство (4) можно интерпретировать следующим образом: эффект от воздействия двух факторов равен сумме эффектов от изолированного воздействия каждого в отдельности при условии, что значение уровня каждого фактора равно 1 при воздействии этого фактора и равно 0 при отсутствии его воздействия.

### 3. Система равенства рисков для трех факторов

Пусть  $A, B, C$  – двухуровневые факторы. Тогда имеется 8 ячеек со значением отклика (случайной величины с двумя уровнями 0,1) при данном наборе факторов  $(0,0,0), (0,0,1), (0,1,0), (0,1,1), (1,0,0), (1,0,1), (1,1,0), (1,1,1)$ . Математическое ожидание (теоретическое среднее значение) отклика в этих ячейках обозначим через  $p_{ijk}$ ,  $i, j, k \in \mathbb{B}$ . Поскольку отклик бинарный и принимает значение из  $\mathbb{B}$ , имеем

$$p_{ijk} = P(Y = 1 | A = i, B = j, C = k), i, j, k \in \mathbb{B}$$

Тогда возможно 12 взаимодействий двух факторов при всевозможных значениях уровней этих факторов при фиксированном значении третьего фактора и 8 взаимодействий трех факторов при всевозможных значениях уровней этих факторов (всего  $12 + 8 = 20$  взаимодействий). Их характеризуют величины  $\gamma_{ij}^{AB}$ ,  $\gamma_{ik}^{AC}$ ,  $\gamma_{jk}^{BC}$  и  $\gamma_{ijk}^{ABC}$ ,  $i, j, k \in \mathbb{B}$  соответственно. Аналогично двухфакторному случаю можно получить следующие выражения для этих взаимодействий и равенство их нулю как выражение отсутствия всяких взаимодействий:

$$\begin{cases} \gamma_{ij}^{AB} = p_{ij\cdot} - p_{i\cdot\cdot} - p_{\cdot j\cdot} + p_{\cdot\cdot\cdot} = 0, i, j \in \mathbb{B} & (5.1) \\ \gamma_{ik}^{AC} = p_{i\cdot k} - p_{i\cdot\cdot} - p_{\cdot\cdot k} + p_{\cdot\cdot\cdot} = 0, i, k \in \mathbb{B} & (5.2) \\ \gamma_{jk}^{BC} = p_{\cdot jk} - p_{\cdot j\cdot} - p_{\cdot\cdot k} + p_{\cdot\cdot\cdot} = 0, j, k \in \mathbb{B} & (5.3) \\ \gamma_{ijk}^{ABC} = p_{ijk} - p_{ij\cdot} - p_{i\cdot k} - p_{\cdot jk} + p_{i\cdot\cdot} + p_{\cdot j\cdot} + p_{\cdot\cdot k} - p_{\cdot\cdot\cdot} = 0, i, j, k \in \mathbb{B}, & (5.4) \end{cases} \quad (5)$$

где, как и выше, точка означает усреднение по соответствующему индексу:

$$p_{ij\cdot} = \frac{1}{2}(p_{ij0} + p_{ij1}), \quad p_{i\cdot\cdot} = \frac{1}{4}(p_{i00} + p_{i01} + p_{i10} + p_{i11}), \quad p_{\cdot\cdot\cdot} = \frac{1}{8} \sum_{i, j, k \in \mathbb{B}} p_{ijk}$$

**Лемма 1.** Пусть  $q_{ij} = p_{ij\cdot}$  для  $i, j \in \mathbb{B}$ , тогда  $q_{i\cdot} = p_{i\cdot\cdot}$ ,  $q_{\cdot j} = p_{\cdot j\cdot}$ ,  $q_{\cdot\cdot} = p_{\cdot\cdot\cdot}$

Утверждения леммы 1 проверяются простыми вычислениями.

**Лемма 2.** Подсистема (5.4) системы (5) равносильна единственному уравнению

$$p_{000} - (p_{001} + p_{010} + p_{100}) + (p_{011} + p_{101} + p_{110}) - p_{111} = 0 \quad (6)$$

**Доказательство** проводится прямыми вычислениями.

Решим систему (5) равенства рисков для трех факторов.

**Теорема 2.** Система (5) из 20 уравнений равносильна системе (7) из 4 уравнений

$$\begin{cases} p_{000} - p_{010} - p_{100} + p_{110} = 0 & (7.1) \\ p_{000} - p_{001} - p_{100} + p_{101} = 0 & (7.2) \\ p_{000} - p_{001} - p_{010} + p_{011} = 0 & (7.3) \\ 2p_{000} - p_{001} - p_{010} - p_{100} + p_{111} = 0 & (7.4) \end{cases} \quad (7)$$

**Доказательство.** Сделав замену  $q_{ij} = p_{ij\cdot}$ , из теоремы 1 и леммы 1 получаем, что подсистема (5.1) системы (5) равносильна уравнению

$$p_{00\cdot} - p_{01\cdot} - p_{10\cdot} + p_{11\cdot} = 0 \quad (8)$$

Аналогично получаем уравнения

$$p_{0\cdot 0} - p_{0\cdot 1} - p_{1\cdot 0} + p_{1\cdot 1} = 0 \quad (9)$$

$$p_{\cdot 00} - p_{\cdot 01} - p_{\cdot 10} + p_{\cdot 11} = 0 \quad (10)$$

Таким образом, используя лемму 2, получаем, что система (5) равносильна системе из уравнений (6), (8), (9), (10).

Прибавим к левой и правой частям уравнения (6), соответственно, левую и правую части уравнения (8), умноженные на 2, и затем разделим каждую часть полученного уравнения на два, тогда получим уравнение (7.1). Аналогично получим уравнения (7.2) и (7.3). Подставив в уравнение (6) переменные  $p_{110}$ ,  $p_{101}$ ,  $p_{011}$ , выраженные из уравнений (7.1), (7.2), (7.3), получаем уравнение (7.4) и отсюда имеем равносильность системы уравнений (6), (7.1), (7.2), (7.3) системе (7).

Введем определение, аналогичное определению 1.

**Определение 2.** Пусть значение уровня фактора  $A$  (фактора  $B$ , фактора  $C$ ) равно 1 при воздействии фактора  $A$  (фактора  $B$ , фактора  $C$ ) и равно 0 при отсутствии воздействия этого фактора. И пусть индексы  $i, j, k \in \mathbb{B}$  не равны нулю одновременно. Величину  $\Delta_{ijk} = p_{ijk} - p_{000}$  назовем эффектом от воздействия фактора  $A$  (при отсутствии воздействия фактора  $A$ ), если  $i = 1$  ( $i = 0$ ), воздействия фактора  $B$  (при отсутствии воздействия фактора  $B$ ), если  $j = 1$  ( $j = 0$ ), воздействия фактора  $C$  (при отсутствии воздействия фактора  $C$ ), если, соответственно,  $k = 1$  ( $k = 0$ ). Полагаем по определению, что  $\Delta_{000} = 0$ .

**Пример 2.**  $\Delta_{101}$  – эффект от воздействия факторов  $A$  и  $C$  при отсутствии воздействия фактора  $B$ .

Из теоремы 2 и определения величин  $\Delta_{ijk}$ ,  $i, j, k \in \mathbb{B}$ , вытекает

**Следствие 2.** Система (5) равенства рисков для трех факторов равносильна системе

$$\begin{cases} \Delta_{110} = \Delta_{010} + \Delta_{100}, & (11.1) \\ \Delta_{101} = \Delta_{001} + \Delta_{100}, & (11.2) \\ \Delta_{011} = \Delta_{001} + \Delta_{010}, & (11.3) \\ \Delta_{111} = \Delta_{001} + \Delta_{010} + \Delta_{100} & (11.4) \end{cases} \quad (11)$$

Интерпретация этого результата состоит в следующем: (а) эффект от воздействия трех факторов равен сумме эффектов от изолированного воздействия каждого в отдельности; (б) эффект от воздействия любых двух факторов равен сумме эффектов от изолированного воздействия каждого в отдельности при отсутствии воздействия третьего. При этом предполагается, что значение уровня каждого фактора равно 1 при воздействии этого фактора и равно 0 при отсутствии этого воздействия.

#### 4. Детерминистическая модель для двух и трех факторов

Рассмотрим детерминистическую модель [7] для двух факторов, где  $p_{ij} \in \mathbb{B}$  для любых  $i, j \in \mathbb{B}$ . Тогда величину  $p_{ij} \in \mathbb{B}$  можно рассматривать как значение булевой функции  $f \in \mathbb{B}(x_1, x_2)$  от значений уровней факторов  $x_1 = i$ ,  $x_2 = j$ :  $p_{ij} = f(i, j)$ . Здесь  $\mathbb{B}(x_1, x_2)$  – множество всех булевых функций от переменных  $x_1, x_2$ . Булева функция  $f$  тогда представляет отклик, а переменные  $x_1, x_2$  – факторы. Для дальнейшего исследования удобно рассматривать булевы функции, представленные в виде ДНФ (дизъюнктивной нормальной формы).

**Теорема 3.** Пусть для некоторой функции  $f \in \mathbb{B}(x_1, x_2)$  имеет место равенство  $p_{ij} = f(i, j)$ ,  $i, j \in \mathbb{B}$ , и для величин  $p_{ij}$  справедлива система (2). Тогда  $f \in \{0, 1, x_1, x_2, \bar{x}_1, \bar{x}_2\}$ .

**Доказательство.** Пусть  $f$  – ненулевая булева функция из  $\mathbb{B}(x_1, x_2)$ . По теореме 1 система (2) равносильна уравнению (3) и уравнению (4). Пусть  $p_{11} = f(1, 1) = 1$ . Если  $p_{00} = f(0, 0) = 1$ , то из уравнений (3) или (4) легко следует, что  $p_{01} = p_{10} = f(0, 1) = f(1, 0) = 1$ , т. е.  $f = 1$ . Если  $f(0, 0) = 0$ , то из этих же уравнений  $f(1, 0) = 1$ ,  $f(0, 1) = 0$  или  $f(1, 0) = 0$ ,  $f(0, 1) = 1$ , т. е.  $f = x_1$  или  $f = x_2$  соответственно. В случае  $p_{11} = f(1, 1) = 0$  аналогично получаем, что  $f = \bar{x}_1$ , или  $f = \bar{x}_2$ .

Рассмотрим детерминистическую модель для трех факторов, где  $p_{ijk} \in \mathbb{B}$  для любых  $i, j, k \in \mathbb{B}$ . Аналогично случаю двух факторов величину  $p_{ijk} \in \mathbb{B}$  можно рассмат-

ривать как значение булевой функции  $f \in \mathbb{B}(x_1, x_2, x_3)$  от значений уровней факторов  $x_1 = i, x_2 = j, x_3 = k: p_{ijk} = f(i, j, k)$ . Здесь  $\mathbb{B}(x_1, x_2, x_3)$  – множество всех булевых функций от переменных  $x_1, x_2, x_3$ . Для трех факторов справедливо утверждение, аналогичное теореме 3.

**Теорема 4.** Пусть для некоторой булевой функции  $f \in \mathbb{B}(x_1, x_2, x_3)$  имеет место равенство  $p_{ijk} = f(i, j, k), i, j, k \in \mathbb{B}$ , и для величин  $p_{ijk}$  справедлива система (5). Тогда  $f \in \{0, 1, x_1, x_2, x_3, \bar{x}_1, \bar{x}_2, \bar{x}_3\}$ .

**Доказательство.** Заметим, что  $\Delta_{ijk} = f(i, j, k) - f(0, 0, 0)$ . Пусть  $f$  – ненулевая булева функция из  $\mathbb{B}(x_1, x_2, x_3)$ , тогда снова сначала рассмотрим ситуацию, когда  $f(1, 1, 1) = 1$ . Возможно два случая:  $f(0, 0, 0) = 1$  и  $f(0, 0, 0) = 0$ . Рассмотрим первый случай, тогда  $\Delta_{111} = 0$  и  $\Delta_{ijk} \leq 0$  для любых  $i, j, k \in \mathbb{B}$ . Следовательно, из уравнения (11.4) имеем  $\Delta_{100} = \Delta_{010} = \Delta_{001} = 0$ . Тогда из уравнений (11.1)–(11.3) получаем равенства  $\Delta_{110} = \Delta_{101} = \Delta_{011} = 0$ , а это и значит, что  $f = 1$ . Рассмотрим второй случай, тогда  $\Delta_{111} = 1$  и  $\Delta_{ijk} \geq 0$  для любых  $i, j, k \in \mathbb{B}$ . Снова применяем уравнение (11.4): ровно одна из величин  $\Delta_{100}, \Delta_{010}, \Delta_{001}$  равна 1, пусть без ограничения общности это будет  $\Delta_{100}$ , т. е.  $\Delta_{100} = 1, \Delta_{010} = 0, \Delta_{001} = 0$ . Применяем уравнения (11.1)–(11.3) и получаем:  $\Delta_{110} = \Delta_{101} = 1, \Delta_{011} = 0$ . Таким образом,  $f(1, j, k) = 1, f(0, j, k) = 0$  для всех  $j, k \in \mathbb{B}$ , а это значит, что  $f = x_1$ . Таким образом, при условии  $f(1, 1, 1) = 1$  имеем  $f \in \{1, x_1, x_2, x_3\}$ . При условии  $f(1, 1, 1) = 0$  совершенно аналогично можно получить, что  $f \in \{\bar{x}_1, \bar{x}_2, \bar{x}_3\}$ .

**Следствие 3.** В случае детерминистической модели для двух (трех) факторов система (2) (система (5)) равенства рисков выполняется для тех и только тех булевых функций, которые существенно зависят не более чем от одной переменной.

**Пример 3.** При  $n = 2$  система (2) равенства рисков выполняется только для функций

$$f \in \{0, 1, x_1, \bar{x}_1, x_2, \bar{x}_2\}$$

и не выполняется для всех остальных функций  $f \in \{x_1 \vee x_2, \bar{x}_1 \vee x_2, x_1 \vee \bar{x}_2, \bar{x}_1 \vee \bar{x}_2, x_1 x_2, \bar{x}_1 x_2, x_1 \bar{x}_2, \bar{x}_1 \bar{x}_2, x_1 x_2 \vee \bar{x}_1 \bar{x}_2, \bar{x}_1 x_2 \vee x_1 \bar{x}_2\}$ .

**Пример 4.** При  $n = 3$  система равенства рисков (5) выполняется только для функций  $f \in \{0, 1, x_1, \bar{x}_1, x_2, \bar{x}_2, x_3, \bar{x}_3\}$  и не выполняется, например, для функций

$$f \in \{x_1 \vee x_2, x_1 \vee x_2 \vee x_3, x_1 x_2, x_1 x_2 x_3, x_1 \bar{x}_2 \vee \bar{x}_1 x_2, x_1 x_2 \vee x_3\}$$

## 5. Совместное действие бинарных факторов в бинарном отклике

Следствие 3 означает, что если считать, что отсутствие взаимодействия данных факторов в булевой функции от двух (трех) переменных выражается системой равенства рисков (2) (системой (5)), то любая булева функция, существенно зависящая более чем от одной переменной, представляет взаимодействие факторов.

В работах [8–11] была развита теория совместного действия бинарных факторов в бинарном отклике в рамках концепции бинарной теории достаточных причин, применяемой в токсикологии, эпидемиологии и доказательной медицине. Значение отклика, равное 1, в этих приложениях означает обычно заболевание или летальный исход. В частности, в этих работах было дано более тонкое определение взаимодействия бинарных факторов (переменных) в бинарном отклике (булевой функции). Приведем это определение.

**Определение 3** [8–10]. Будем говорить, что в булевой функции  $f \in \mathbb{B}(x_1, x_2)$  ( $f \in \mathbb{B}(x_1, x_2, x_3)$ ) имеет место *совместное действие* двух (трех) переменных  $x = x_1, x_2$  ( $x = x_1, x_2, x_3$ ), если существует такой набор  $\alpha = (\alpha_1, \alpha_2) \in \mathbb{B}^2$  ( $\alpha = (\alpha_1, \alpha_2, \alpha_3) \in \mathbb{B}^3$ ), что конъюнкция  $x^\alpha = x_1^{\alpha_1}x_2^{\alpha_2}$  ( $x^\alpha = x_1^{\alpha_1}x_2^{\alpha_2}x_3^{\alpha_3}$ ) является простым импликантом [12–14] булевой функции  $f$ . Будем говорить, что в этом случае совместное действие в булевой функции достигается для  $x = \alpha$ .

Здесь для  $\gamma \in \mathbb{B}$  символ  $x^\gamma$  обозначает  $x$ , если  $\gamma = 1$ , и  $\bar{x}$ , если  $\gamma = 0$ .

**Пример 5.** При  $n = 2$  совместное действие переменных  $x_1, x_2$  есть в булевых функциях

$$f \in \{x_1x_2, \bar{x}_1x_2, x_1\bar{x}_2, \bar{x}_1\bar{x}_2, x_1x_2 \vee \bar{x}_1\bar{x}_2, \bar{x}_1x_2 \vee x_1\bar{x}_2\}$$

и нет во всех остальных булевых функциях

$$f \in \{0, 1, x_1, \bar{x}_1, x_2, \bar{x}_2, x_1 \vee x_2, \bar{x}_1 \vee x_2, x_1 \vee \bar{x}_2, \bar{x}_1 \vee \bar{x}_2\}$$

**Пример 6.** При  $n = 3$  совместное действие переменных  $x_1, x_2, x_3$  есть, например, в булевых функциях  $f \in \{x_1x_2x_3, \bar{x}_1\bar{x}_2\bar{x}_3 \vee x_1x_2, \bar{x}_1\bar{x}_2\bar{x}_3 \vee x_1x_2x_3\}$  и нет его, например, в булевых функциях

$$f \in \{0, 1, x_1, \bar{x}_1, x_1x_2, x_1 \vee x_3, \bar{x}_1 \vee x_2 \vee \bar{x}_3, x_1x_3 \vee x_2x_3\}$$

Из следствия 3 и определения 3 непосредственно вытекает

**Следствие 4.** Если в булевой функции  $f \in \mathbb{B}(x_1, x_2)$  ( $f \in \mathbb{B}(x_1, x_2, x_3)$ ) есть совместное действие переменных  $x_1, x_2$  ( $x_1, x_2, x_3$ ), то для  $f$  не выполняется система (2) (система (5)) равенства рисков, где  $p_{ij} = f(i, j)$ ,  $i, j \in \mathbb{B}$  ( $p_{ijk} = f(i, j, k)$ ,  $i, j, k \in \mathbb{B}$ ).

Обратное неверно, а именно: 1) для  $n = 2$  в булевых функциях вида  $f = x_1^{\alpha_1} \vee x_2^{\alpha_2}$ , где  $\alpha_1, \alpha_2 \in \mathbb{B}$ , по следствию 3 не выполняется система (2) равенства рисков, но по определению 3 нет совместного действия переменных  $x_1, x_2$ ; 2) для  $n = 3$  в булевых функциях вида  $f = x_1^{\alpha_1} \vee x_2^{\alpha_2} \vee x_3^{\alpha_3}$ ,  $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{B}$ , аналогично не выполняется система (5) равенства рисков, но по определению 3 нет совместного действия переменных  $x_1, x_2, x_3$ . Таким образом, в рамках детерминистической модели определение 3 совместного действия бинарных факторов более точно характеризует взаимодействие бинарных факторов, чем традиционное определение через невыполнение системы равенства рисков.

Рассмотрим примеры откликов (булевых функций) для двух факторов (двух переменных) и взаимодействие этих факторов в них.

**Пример 7.** Известно, что риск развития рака легких среди курящих значительно повышается в присутствии амфиболового асбеста [9, 15]. В этом случае бинарным фактором  $x_1$  можно считать курение, точнее,  $x_1 = 1$  для курящего человека и  $x_1 = 0$  для некурящего, бинарным фактором  $x_2$  – присутствие амфиболового асбеста ( $x_2 = 1$  – амфиболовый асбест присутствует,  $x_2 = 0$  – отсутствует), а булевой функцией, характеризующей отклик, функцию  $f = x_1x_2$  – развитие рака. По определению 3 в  $f$  есть совместное действие факторов  $x_1, x_2$  при уровнях значений факторов  $x_1 = x_2 = 1$ , а значит, по следствию 3 не выполняется система (2) равенства рисков, где  $p_{ij} = f(i, j)$ ,  $i, j \in \mathbb{B}$ .



**Пример 8.** При одном из обследований несовершеннолетних пациентов выявилась ситуация, когда у детей, проживавших в местности, где произошел выброс соединений и свинца, и кадмия, результаты анализов были практически такими же, как и у детей, проживавших в местности, где не было выброса ни того, ни другого. В то же время, у детей, проживавших в местности, где были выбросы только соединений свинца или только соединений кадмия, результаты анализов были плохими. Интересно, что сами по себе соединения свинца и кадмия не нейтрализуют друг друга. В данном случае бинарным фактором  $x_1$  можно считать присутствие выбросов соединений свинца, а фактором  $x_2$  – присутствие соединений кадмия. Булевой функцией, характеризующей отклик, является функция  $f = \bar{x}_1 x_2 \vee x_1 \bar{x}_2$ . В  $f$  по определению 3 есть совместное действие факторов  $x_1, x_2$  при уровнях значений факторов  $x_1 = 0, x_2 = 1$  (проживание в местности, с выбросами только соединений кадмия) или при  $x_1 = 1, x_2 = 0$  (проживание в местности, с выбросами только соединений свинца), и не выполняется система (2) равенства рисков.

**Пример 9.** Известно, что цианиды в промышленных отходах токсичны для аквафауны. Но если в этих отходах присутствуют соединения никеля, которые сами по себе тоже является токсичным, то ионы никеля могут образовывать с цианид-ионами нетоксичные соединения [8]. Здесь, как и в предыдущем примере, булева функция, характеризующая отравление аквафауны, зависящая от факторов  $x_1$  – присутствия цианидов и  $x_2$  – присутствия никеля, равна  $f = \bar{x}_1 x_2 \vee x_1 \bar{x}_2$ . И так же, как и в предыдущем примере, мы имеем совместное действие факторов  $x_1, x_2$  при уровнях значений факторов  $x_1 = 0, x_2 = 1$  (присутствие только никеля) или при  $x_1 = 1, x_2 = 0$  (присутствие только цианидов) и невыполнение системы (2) равенства рисков.

Булева функция следующего примера может возникнуть при описании любых двух несовместимых лекарственных препаратов.

**Пример 10.** При лечении заболеваний щитовидной железы могут назначать высокие дозы препарата Йодомарина, действующим веществом которого является йодид калия. Если при этом для лечения артериальной гипертензии назначаются калийсберегающие диуретики, то может произойти гиперкалиемия, что значительно ухудшит состояние организма. Таким образом, булева функция, означающая в данном случае ослабление общего состояния здоровья, для пациентов с заболеванием щитовидной железы и артериальной гипертензией равна  $f = \bar{x}_1 \bar{x}_2 \vee x_1 x_2$ , где фактор  $x_1$  – прием высоких доз препарата Йодомарина (Й),  $x_2$  – прием калийсберегающих диуретиков (КД). В функции  $f$  есть совместное действие факторов  $x_1, x_2$  при уровнях значений факторов  $x_1 = 0, x_2 = 0$  (отсутствие приема высоких доз Й и приема КД) или при  $x_1 = 1, x_2 = 1$  (одновременный прием двух препаратов), и, конечно, не выполняется система (2) равенства рисков.

**Пример 11.** Если рассматривать факторы  $x_1, x_2$  как курение и злоупотребление пациентами спиртных напитков соответственно, а откликом считать булеву функцию  $f$ , означающую развитие у курящих пациентов или пациентов, злоупотребляющих спиртными напитками, сердечно-сосудистых заболеваний, то  $f = x_1 \vee x_2$ , поскольку это заболевание возникает при присутствии хотя бы одного из этих двух факторов. В этой булевой функции нет совместного действия переменных  $x_1, x_2$  по определению 3, хотя по следствию 4 система (2) равенства рисков не выполняется.

Теперь рассмотрим примеры откликов в случае трех факторов и взаимодействие этих факторов в них.

**Пример 12.** Пусть фактор  $x_1$  – низкая усваиваемость железа, фактор  $x_2$  – аллергия на препараты железа, фактор  $x_3$  – низкая свертываемость крови, а  $f$  – булева функция, означающая присутствие у пациентов анемии. Тогда  $f = x_1 x_2 x_3$ . Дей-

ствительно, это заболевание проявляется только при присутствии всех этих трех факторов. В этой булевой функции есть совместное действия переменных  $x_1, x_2, x_3$  по определению 3, и по следствию 4 не выполняется система равенства рисков (5).

В работах [9, 11] введено и изучено понятие совместного действия двух бинарных факторов в отклике, зависящем от трех факторов.

**Определение 4.** В булевой функции (отклике)  $f \in \mathbb{B}(x_1, x_2, x_3)$  имеется совместное действие двух булевых переменных (факторов)  $x_i = (x_j)_{j \in I}$  для двухэлементного упорядоченного подмножества  $I$  множества  $\{1, 2, 3\}$ , если в этой функции (отклике) есть совместное действие этих двух переменных при фиксированном значении третьей переменной  $x_j = \beta, j \notin I, \beta \in \mathbb{B}$ .

В этих же работах показано, что если в булевой функции есть совместное действие трех переменных, то в ней есть и совместное действие любых двух переменных.

**Пример 13.** Во всех булевых функциях из примера 6, где есть совместное действие трех переменных, есть и совместное действие любых двух переменных. Кроме того, в следующих булевых функциях из примера 6 по определению 4 есть совместное действие двух переменных, но по определению 3 нет совместного действия трех переменных: в  $f = x_1 x_2$  есть совместное действие переменных  $x_1, x_2$  для значений уровней  $x_1 = 1$  и  $x_2 = 1$  при фиксированном значении уровня  $x_3 = 0$  или  $x_3 = 1$ ; в  $f = x_1 x_3 \vee x_2 x_3$  есть совместное действие переменных  $x_1, x_3$  для значений уровней  $x_1 = 1$  и  $x_3 = 1$  при фиксированном значении уровня  $x_2 = 0$  или совместное действие переменных  $x_2, x_3$  для значений уровней  $x_2 = 1$  и  $x_3 = 1$  при фиксированном значении уровня  $x_1 = 0$ .

**Пример 14.** Если рассматривать зависимость отклика  $f$  – тяжелого течения болезни COVID-19 от факторов  $x_1$  – сахарного диабета,  $x_2$  – возраста выше 65 лет и  $x_3$  – вакцинации от COVID-19, то легко понять, что  $f = x_1 \bar{x}_3 \vee x_2 \bar{x}_3$ . В этой булевой функции нет совместного действия трех переменных  $x_1, x_2, x_3$  по определению 3, но по определению 4 есть совместное действие переменных  $x_1, x_3$  для значений  $x_1 = 1$  (присутствие сахарного диабета) и  $x_3 = 0$  (отсутствие прививки) при фиксированном значении переменной  $x_2 = 0$  (возраст не выше 65), а также есть совместное действие переменных  $x_2, x_3$  для значений  $x_2 = 1$  (возраст выше 65) и  $x_3 = 0$  (отсутствие прививки) при фиксированном значении переменной  $x_1 = 0$  (отсутствии сахарного диабета). Кроме того, по следствию 4 не выполняется система (5) равенства рисков.

Введение и исследование понятия совместного действия двух факторов в отклике, зависящем от трех факторов, – это еще одно обстоятельство, подчеркивающее, что определение совместного действия бинарных факторов в бинарном отклике, развитое в работах [8–11], в рамках детерминистической модели более точно описывает взаимодействие факторов, чем просто невыполнение системы равенства рисков, соответствующей данному отклику.

## 6. Инвариантность системы равенства рисков относительно группы симметрий эксперимента

При исследовании совместного действия двух (трех) бинарных факторов в отклике неважно, как кодируются уровни факторов. Так возникает группа симметрий эксперимента, являющаяся группой  $G_2$  ( $G_3$ ) автоморфизмов булева квадрата  $\mathbb{B}^2$  (булева куба  $\mathbb{B}^3$ ) как графа, две вершины которого соединены ребром тогда и только тогда, когда расстояние Хэмминга между ними равно 1 [8, 16, 17]. Автоморфизмы булева квадрата  $\mathbb{B}^2$  (булева куба  $\mathbb{B}^3$ ) естественным образом продолжаются на булеву

алгебру  $\mathbb{B}(x_1, x_2)$  ( $\mathbb{B}(x_1, x_2, x_3)$ ). Действие группы  $G_2$  ( $G_3$ ) на булевой алгебре  $\mathbb{B}(x_1, x_2)$  ( $\mathbb{B}(x_1, x_2, x_3)$ ) заключается в следующих последовательно выполненных операциях: любой перестановке переменных в каждой функции и навешивании отрицаний на некоторые переменные. Группа  $\mathbb{B}(x_1, x_2)$  ( $\mathbb{B}(x_1, x_2, x_3)$ ) изоморфна группе симметрий квадрата (куба) [17, 18]. Действие группы  $G_2$  ( $G_3$ ) на алгебре  $\mathbb{B}(x_1, x_2)$  ( $\mathbb{B}(x_1, x_2, x_3)$ ) порождает естественное разбиение этой алгебры на орбиты действия этой группы – типы совместного действия факторов [8, 16–18]. Орбиты булевой функции  $f$  относительно этого действия будем обозначать через  $\langle f \rangle$ . Совместное действие двух (трех) факторов инвариантно относительно действия группы  $G_2$  ( $G_3$ ) [8, 9, 18].

Группа  $G_2$  ( $G_3$ ), действуя на булевом квадрате  $\mathbb{B}^2$  (на булевом кубе  $\mathbb{B}^3$ ), действует и на множестве переменных  $P_2 = \{p_{ij} \mid i, j \in \mathbb{B}\}$  ( $P_3 = \{p_{ijk} \mid i, j, k \in \mathbb{B}\}$ ). Поэтому группу  $G_2$  ( $G_3$ ) можно считать подгруппой группы  $Sym(P_2)$  ( $Sym(P_3)$ ) перестановок множества переменных  $P_2$  ( $P_3$ ). Хорошо известно представление группы  $Sym(P_2)$  ( $Sym(P_3)$ ) перестановочными матрицами [19]. Таким образом, группа  $Sym(P_2)$  ( $Sym(P_3)$ ) изоморфным образом вкладывается в группу  $GL(2^2, \mathbb{Q})$  ( $GL(2^3, \mathbb{Q})$ ) всех невырожденных матриц четвертого (восьмого) порядка над полем  $\mathbb{Q}$  рациональных чисел или, что то же самое, в группу всех обратимых линейных преобразований векторного пространства  $\mathbb{Q}^{2^2}$  ( $\mathbb{Q}^{2^3}$ ) над полем  $\mathbb{Q}$  [20]. Кроме того, поскольку векторное пространство  $\mathbb{Q}^{2^2}$  ( $\mathbb{Q}^{2^3}$ ) изоморфно векторному пространству  $\mathbb{Q}[P_2]$  ( $\mathbb{Q}[P_3]$ ) всех многочленов с коэффициентами из  $\mathbb{Q}$  от переменных из  $G_2$  ( $G_3$ ), группа  $G_2$  ( $G_3$ ) действует также и на  $\mathbb{Q}[P_2]$  ( $\mathbb{Q}[P_3]$ ).

**Теорема 5.** Система (2) равенства рисков (система (5) равенства рисков) инвариантна относительно действия группы  $G_2$  ( $G_3$ ).

**Доказательство.** Легко понять, что система (2) инвариантна относительно действия группы  $G_2$ . Аналогично каждая из подсистем (5.1), (5.2), (5.3), (5.4) инвариантна относительно действия группы  $G_3$ , а значит, и система (5) инвариантна относительно действия этой группы.

Таким образом, о выполнении или невыполнении системы равенства рисков можно говорить не только для данной булевой функции  $f$ , но для всего класса  $\langle f \rangle$ .

Рассмотрим обобщение определений 1, 2.

**Определение 5.** Пусть значение уровня фактора  $A$  (фактора  $B$ ) равно  $\alpha_1 \in \mathbb{B}$  ( $\alpha_2 \in \mathbb{B}$ ) при воздействии фактора  $A$  (фактора  $B$ ) и равно  $\bar{\alpha}_1$  ( $\bar{\alpha}_2$ ) при отсутствии воздействия этого фактора. Пусть индексы  $i, j \in \mathbb{B}$  не равны нулю одновременно. Величину  $\Delta_{\alpha, ij} = p_{i \oplus \alpha_1, j \oplus \alpha_2} - p_{\alpha_1, \alpha_2}$ , где  $\alpha = (\alpha_1, \alpha_2) \in \mathbb{B}^2$  и  $\oplus$  – сложение по модулю 2, назовем эффектом от воздействия фактора  $A$  (при отсутствии воздействия фактора  $A$ ), если  $i = 1$  ( $i = 0$ ), и воздействия фактора  $B$  (при отсутствии воздействия фактора  $B$ ), если  $j = 1$  ( $j = 0$ ). Полагаем по определению, что  $\Delta_{\alpha, 00}$ .

Обобщением следствия 1 является утверждение

**Следствие 5.** Для любого набора  $\alpha = (\alpha_1, \alpha_2) \in \mathbb{B}^2$  система (2) равенства рисков равносильна уравнению

$$\Delta_{\alpha, 11} = \Delta_{\alpha, 01} + \Delta_{\alpha, 10}. \quad (12)$$

**Доказательство.** Очевидно, многочлен  $\Delta_{\alpha,ij} \in \mathbb{Q}[P_2]$  является образом многочлена  $\Delta_{ij} \in \mathbb{Q}[P_2]$  при действии сдвига  $t_\alpha(x) = x \oplus \alpha$  на множестве  $\mathbb{Q}[P_2]$ . Учитывая, что  $t_\alpha(x) \in G_2$ , по следствию 1 и теореме 5 получаем требуемое.

**Пример 15.** Пусть значения уровней факторов  $A, B$  равны  $\alpha_1 = 1$  и  $\alpha_2 = 0$ , соответственно, при воздействии этих факторов. Тогда по следствию 5 система равенства рисков (2) равносильна равенству (12):

$$p_{1\oplus 1,1\oplus 0} - p_{10} = (p_{0\oplus 1,1\oplus 0} - p_{10}) + (p_{1\oplus 1,0\oplus 0} - p_{10}),$$

т. е. равенству  $p_{01} - p_{10} = (p_{11} - p_{10}) + (p_{00} - p_{10})$ .

Следствие 5 можно интерпретировать следующим образом: эффект от воздействия двух факторов равен сумме эффектов от изолированного воздействия каждого в отдельности независимо от значений уровней факторов, при котором происходит воздействие этих факторов.

Аналогично можно ввести величины  $\Delta_{\alpha,ijk} = p_{i\oplus \alpha_1, j\oplus \alpha_2, k\oplus \alpha_3} - p_{\alpha_1, \alpha_2, \alpha_3}$  для  $\alpha = (\alpha_1, \alpha_2, \alpha_3) \in \mathbb{B}^3$ , если индексы  $i, j, k \in \mathbb{B}$  не равны нулю одновременно, принимая, что  $\Delta_{\alpha,000} = 0$ , и доказать аналог следствия 5.

**Следствие 6.** Для любого набора  $\alpha = (\alpha_1, \alpha_2, \alpha_3) \in \mathbb{B}^3$  система (5) равенства рисков равносильна системе

$$\begin{cases} \Delta_{\alpha,110} = \Delta_{\alpha,010} + \Delta_{\alpha,100} \\ \Delta_{\alpha,101} = \Delta_{\alpha,001} + \Delta_{\alpha,100} \\ \Delta_{\alpha,011} = \Delta_{\alpha,001} + \Delta_{\alpha,010} \\ \Delta_{\alpha,111} = \Delta_{\alpha,001} + \Delta_{\alpha,010} + \Delta_{\alpha,100}. \end{cases}$$

Интерпретировать этот результат можно следующим образом: (а) эффект от воздействия трех факторов равен сумме эффектов от изолированного воздействия каждого в отдельности; (б) эффект от воздействия любых двух факторов равен сумме эффектов от изолированного воздействия каждого в отдельности при отсутствии воздействия третьего. И этот результат не зависит от значений уровней факторов, при котором происходит воздействие этих факторов.

## 7. Заключение

В настоящей работе:

1. получена и решена система равенства рисков для двух и трех бинарных факторов в случае бинарного отклика;
2. показано, что система равенства рисков для двух и трех факторов инвариантна относительно действия группы симметрий эксперимента;
3. получена интерпретация решения системы равенства рисков для двух факторов: эффект от воздействия двух факторов равен сумме эффектов от изолированного воздействия каждого в отдельности независимо от значений уровней факторов, при котором происходит воздействие этих факторов;
4. получена интерпретация решения системы равенства рисков для трех факторов: (а) эффект от воздействия трех факторов равен сумме эффектов от изолированного воздействия каждого в отдельности; (б) эффект от воздействия

любых двух факторов равен сумме эффектов от изолированного воздействия каждого в отдельности при отсутствии воздействия третьего; этот результат не зависит от значений уровней факторов, при котором происходит воздействие этих факторов;

5. показано, что в рамках детерминистической модели и для двух, и для трех факторов система равенства рисков выполняется только для откликов, существенно зависящих не более чем от одного фактора;
6. показано, что в рамках детерминистической модели определение совместного действия двух или трех факторов в бинарном отклике, данное в работах [8–11], более полно описывает взаимодействие этих факторов, чем невыполнение системы равенства рисков.

## 8. Список литературы

1. *Koopman, J. S.* Interaction between discrete causes / J. S. Koopman // *Am J Epidemiol.* – 1981. – Vol. 113, № 6. – С. 716–724. – DOI: 10.1093/oxfordjournals.aje.a113153.
2. *Фишер, Р. А.* Статистические методы для исследователей / Р. А. Фишер. – М. : Госстатиздат, 1958. – 268 с.
3. *Кендалл, М.* Многомерный статистический анализ и временные ряды / М. Кендалл, А. Стьюарт. – М. : Наука, 1976. – 737 с.
4. *Шеффе, Г.* Дисперсионный анализ / Г. Шеффе. – М.: Наука, 1980. – 511 с.
5. *Rothman, K. J.* Modern epidemiology: Third edition / K. J. Rothman, S. Greenland, T. L. Lash // *Modern Epidemiology.* – 2011. – P. 1–758. – EDN YDXWPS.
6. *Szklo, M.* Epidemiology: Beyond the Basics / M. Szklo, F. J. Nieto. – 2<sup>nd</sup> ed. Boston, MA: Jones and Bartlee Publishers, 2007. – 578 с. – ISBN-128411659X.
7. *Miettinen, O. S.* Causal and preventive interdependence. Elementary principles / O. S. Miettinen // *Scand J. Work Environ Health.* – 1982. – Vol. 8, № 3. – С. 159–168. – DOI: 10.5271/sjweh.2479.
8. *Nagrebetskaya, J. V.* Joint action of binary factors in the sufficient causes theory and its classification / J. V. Nagrebetskaya, V. G. Panov // *International J. of Innovative Technology and Exploring Engineering.* – 2019. – Vol. 9, No. 1. – P. 2146–2153. – DOI 10.35940/ijtee.A4702.119119. – EDN LQILAE.
9. *Нагребецкая, Ю. В.* Булева модель частичного взаимодействия бинарных факторов / Ю. В. Нагребецкая, В. Г. Панов // *Вестник Омского университета.* – 2021. – Т. 26, № 3. – С. 10–19. – DOI 10.24147/1812-3996.2021.26(3).10-19. – EDN XVLXOK.
10. *Нагребецкая, Ю. В.* Степень взаимодействия бинарных факторов в теории достаточных причин / Ю. В. Нагребецкая, В. Г. Панов // *Системный анализ в медицине : Материалы XIII Международной научной конференции, Благовещенск, 26–27 сентября 2019 г.* – Благовещенск: Дальневосточный научный центр физиологии и патологии дыхания, 2019. – С. 31–34. – EDN ETTKWN.
11. *Нагребецкая, Ю. В.* Обобщение понятия взаимодействия n факторов в теории достаточных причин и его свойства / Ю. В. Нагребецкая, В. Г. Панов // *Системный анализ в медицине : Материалы XIII Международной научной конференции, Благовещенск, 26–27 сентября 2019 г.* – Благовещенск : Дальневосточный научный центр физиологии и патологии дыхания, 2019. – С. 35–38. – EDN SMROHS.
12. *Яблонский, С. В.* Введение в дискретную математику : учеб. пособие для студентов вузов, обучающихся по специальности «Прикладная математика» / С. В. Яблон-

- ский. – 4 изд., стер. – М. : Высшая школа, 2003. – 384 с. – (Высшая математика). – ISBN 5-06-004681-8. – EDN QJLZCB.
13. *Лидл, Р.* Прикладная абстрактная алгебра: учеб. пособие / Р. Лидл, Г. Пильц; пер. с англ. – Екатеринбург : Изд-во Урал. ун-та, 1996. – 744 с.
  14. *Crama, Y.* Boolean functions. Theory, Algorithms, and Applications / Y. Crama, P. L. Hammer // Cambridge: Cambridge University Press, 2011. – 711 с.
  15. *Boffetta, P.* Occupational exposure to asbestos and man-made vitreous fibres and risk of lung cancer: A multicentre case-control study in Europe / P. Boffetta, R. Carel, A. C. Olsson [et al.] // Occupational and Environmental Medicine. – 2007. – Vol. 64, No. 8. – P. 502–508. – DOI 10.1136/oem.2006.027748. – EDN LKNXVV.
  16. *Panov, V. G.* Boolean algebras and classification of interactions in sufficient-component cause model / V. G. Panov, J. V. Nagrebetskaya // International Journal of Pure and Applied Mathematics. – 2015. – Vol. 98, No. 2. – P. 239–259. – DOI 10.12732/ijpam.v98i2.7. – EDN UENIDL.
  17. *Panov, V. G.* Classification of combined action of binary factors and Coxeter groups / V. G. Panov, J. V. Nagrebetskaya // J. of Discrete Mathematical Sciences and Cryptography. – 2018. – Vol. 21, No. 3. – P. 661–677. – DOI 10.1080/09720529.2016.1222733. – EDN XXXMFV.
  18. *Панов, В. Г.* Алгебраическая трактовка двухфакторной бинарной теории достаточных причин / В. Г. Панов, Ю. В. Нагребецкая // Труды СПИИРАН. – 2013. – № 3(26). – С. 277–296. – EDN QIXKOD.
  19. *Супруненко, Д. А.* Группы матриц / Д. А. Супруненко. – М. : Наука, 1972. – 352 с.
  20. *Винберг, Э. Б.* Курс алгебры / Э. Б. Винберг. – М. : МЦНМО, 2011. – 592 с. ISBN 978-5-94057-685-3. – EDN SDSFZZ.

### Сведения об авторах:

**Нагребецкая Юлия Вацлавовна**, к. ф.-м. н., доцент, Институт естественных наук и математики, департамент математики, механики и компьютерных наук УрФУ, ул. Тургенева, 4, г. Екатеринбург, Россия. Эл. почта: I.V.Nagrebetskaia@urfu.ru.

**Фаткуллина Аэлиа Ильфатовна**, студентка Института естественных наук и математики, департамент математики, механики и компьютерных наук УрФУ.

## JOINT ACTION OF BINARY FACTORS AND RISKS' EQUALITY

J. V. Nagrebeskaya, A. I. Fatkullina

*Ural Federal University named after the first President of Russia B. N. Yeltsin,  
Ekaterinburg, Russia*

*In this paper we study the models of interaction of two and three independent binary factors in the case of binary response, which are widely used in medical and biological sciences. The main attention is paid to the formalization of the concept of "interaction" and its connection with the risks' difference equality. It is shown that this equality can be obtained from equating all interaction terms to zero in their representation used in the analysis of variance. The system of risks' differences equations for two and three factors is obtained and solved, and its invariance with respect to the action of a symmetry group of experiment is shown. The interpretation of the solution for the system of risks' differences equations is presented: in the case of two factors the effect of their impact is equal to the sum of effects of isolated action of each factor separately; in the case of three factors (a) the effect of all these factors is equal to the sum of effects of isolated action of each factor separately; (b) the effect of any two factors is equal to the sum of effects of isolated action of each factor separately in the absence of the third one. In addition, a study of the solution of equality of risks for two and three factors within the framework of the deterministic model, leading to a Boolean formalization of this model, was carried out. It is proved that for the values of the Boolean function representing a given response, the system of risks' differences equations for both two and three factors is satisfied if and only if this Boolean function significantly depends on no more than one variable.*

*It is shown that within the framework of the deterministic model, the concept of "joint action" of binary factors, introduced and studied in the works of the first author, more accurately describes the interaction of these factors than violation of risks' differences equations.*

**Key words:** risks' differences equations; deterministic model; binary theory of sufficient causes; response; joint action of binary factors; Boolean function; automorphism group; group action.

### References

1. *Koopman, J. S. Interaction between discrete causes / J. S. Koopman // Am J Epidem. – 1981. – Vol. 113, № 6. – С. 716–724. – DOI: 10.1093/oxfordjournals.aje.a113153.*
2. *Fisher, R. A. Statistical methods for research workers / R. A. Fisher. – Moskow : Gosstatizdat. – 1958. – 268 p.*
3. *Kendall, M. G. Design and analysis, and time-series / M. G. Kendall, A. Stuart. – Moskow : Nauka. – 1976. – 737 p.*
4. *Scheffe, G. Analysis of variance / G. Scheffe. – Moskow : Nauka, 1980. – 511 p.*

5. *Rothman, K. J.* Modern epidemiology: Third edition / K. J. Rothman, S. Greenland, T. L. Lash // Modern Epidemiology: Third Edition, 2011. – P. 1–758. – EDN YDXWPS.
6. *Szklo, M.* Epidemiology: Beyond the Basics / M. Szklo, F. J. Nieto. – 2<sup>nd</sup> ed. Boston, MA: Jones and Bartlee Publishers, 2007. – 578 с. – ISBN-128411659X.
7. *Miettinen, O. S.* Causal and preventive interdependence. Elementary principles / O. S. Miettinen // Scand J. Work Environ Health. – 1982. – Vol. 8, № 3. – С. 159–168. – DOI: 10.5271/sjweh.2479.
8. *Nagrebetskaya, J. V.* Joint action of binary factors in the sufficient causes theory and its classification / J. V. Nagrebetskaya, V. G. Panov // International J. of Innovative Technology and Exploring Engineering. – 2019. – Vol. 9, No. 1. – P. 2146–2153. – DOI 10.35940/ijitee.A4702.119119. – EDN LQILAE.
9. *Nagrebetskaya, J. V.* Boolean model of partial interaction of binary factors / J. V. Nagrebetskaya, V. G. Panov // Guardian of the Omsk University. – 2021. – Vol. 26, № 3. – P. 10–19. – DOI 10.24147/1812-3996.2021.26(3).10–19.
10. *Nagrebetskaya, J. V.* Interaction degree of binary factors in the sufficient causes theory / J. V. Nagrebetskaya, V. G. Panov // Materials of the XIII International Scientific Conference «System analysis in medicine» (SAM 2019) September 19–20, 2019. Blagoveshchensk. – P. 31–34. – DOI: 10.12737/collection\_5d8335e34b6a76.02467823.
11. *Nagrebetskaya, J. V.* Обобщение понятия взаимодействия  $n$  факторов в теории достаточных причин и его свойства // Materials of the XIII International Scientific Conference «System analysis in medicine» (SAM 2019) September 19–20, 2019. Blagoveshchensk. – P. 35–38. – DOI: 10.12737/collection\_5d8335e34b6a76.02467823.
12. *Yablonskiy, S. V.* Introduction to Discrete Mathematics: Textbook for university students studying in the specialty «Applied Mathematics» / S. V. Yablonskiy. – 4<sup>th</sup> ed. – Moscow: High School, 2003. – 384 p. – (High Mathematics). – ISBN 5-06-004681-8. – EDN QJLZCB.
13. *Lidl, R.* Applied Abstract Algebra: Textbook / R. Lidl, G. Pilz: Translation from English. – Ekaterinburg: Publisher Ural State University, 1996. – 744 p.
14. *Crama, Y.* Boolean functions. Theory, Algorithms, and Applications / Y. Crama, P. L. Hammer // Cambridge: Cambridge University Press, 2011. – 711 с.
15. *Boffetta, P.* Occupational exposure to asbestos and man-made vitreous fibres and risk of lung cancer: A multicentre case-control study in Europe / P. Boffetta, R. Carel, A. C. Olsson [et al.] // Occupational and Environmental Medicine. – 2007. – Vol. 64, No. 8. – P. 502–508. – DOI 10.1136/oem.2006.027748. – EDN LKNXVV.
16. *Panov, V. G.* Boolean algebras and classification of interactions in sufficient-component cause model / V. G. Panov, J. V. Nagrebetskaya // International Journal of Pure and Applied Mathematics. – 2015. – Vol. 98, No. 2. – P. 239–259. – DOI 10.12732/ijpam.v98i2.7. – EDN UENIDL.
17. *Panov, V. G.* Classification of combined action of binary factors and Coxeter groups / V. G. Panov, J. V. Nagrebetskaya // J. of Discrete Mathematical Sciences and Cryptography. – 2018. – Vol. 21, No. 3. – P. 661–677. – DOI 10.1080/09720529.2016.1222733. – EDN XXXMFV.
18. *Panov, V. G.* An Algebraic Interpretation of the Two-Factor Binary Theory of Sufficient Causes / V. G. Панов, J. V. Nagrebetskaya // Proceedings of SPIIRAN. – 2013, № 3(26). – С. 277–296. – EDN QIXKOD.
19. *Suprunenko, D. A.* Groups of Matrices / D. A. Suprunenko. – Moscow : Nauka, 1972. – 352 p.
20. *Vinberg, E. B.* Algebra course / E. B. Vinberg. – Moscow: MCCMO, 2011. – 592 p. – ISBN 978-5-94057-685-3. – EDN SDSFZZ.



# АНАЛИЗ ДИСКРЕТНЫХ ДАННЫХ. ОСНОВНЫЕ ПОНЯТИЯ

В. Г. Панов

*Институт промышленной экологии УрО РАН, г. Екатеринбург, Россия*

*Обзорная статья посвящена методам анализа дискретных данных. Основное внимание уделяется точным методам, основанным на представлении дискретных вероятностных распределений в виде степенных распределений (power series distributions). Рассматриваемые примеры относятся к задачам медицины и биологии.*

**Ключевые слова:** дискретные распределения; степенные распределения; точные методы статистического анализа; статистические гипотезы; доверительные области; мощность критерия; вероятность покрытия.

## 1. Введение

Исследования в медицине [1–5], эпидемиологии [6–10], биологии [11–16], экономике [17–21] и других науках постоянно сталкиваются с необходимостью анализа дискретных данных, для которых, разумеется, разработаны разнообразные методы, однако в большинстве своем они предполагают использование приемов, связанных с большими выборками. Если же выборка существенно ограничена или аппроксимация асимптотическими распределениями недостаточно хорошая, то использование таких методов сомнительно. В этом случае желательно иметь средства точного анализа данных, не обращающиеся к предельным распределениям того или иного типа.

Исторически именно точные методы были первыми, в которых математика применялась для анализа случайных данных, это классическая задача оценки вероятности события при наличии равновероятных исходов и схема повторных независимых испытаний Бернулли [22, 23]. В XX веке актуальность тематики точного анализа дискретных данных была поддержана необходимостью развития непараметрических методов анализа непрерывных данных [24–26]. Однако основное развитие методов теории вероятностей было связано с непрерывными распределениями и предельными теоремами.

Важным событием в возрождении интереса к точным методам и фактически становлением этой тематики как отдельного направления статистических исследований стала почти одновременная публикация работ по точному анализу таблиц сопряженности  $2 \times 2$  в работах Фишера [27, 28], Ирвина [29] и Йейтса [30]. Однако развитие этих идей сдерживалось значительной вычислительной сложностью приложений полученных теоретических результатов. В дальнейшем были разработаны методы, ускоряющие такие вычисления (на основе быстрого преобразования Фурье), а также были разработаны новые алгоритмы таких вычислений. Увеличение

вычислительных возможностей и появление новых алгоритмов сделали методы точного анализа более доступными, что было реализовано в нескольких компьютерных программах (например, StatXact [31], а также отдельные модули точных статистических вычислений в программах SPSS, SAS, R и др.).

Несмотря на значительную потребность в методах точного анализа дискретных данных, изложение темы точных статистических вычислений в учебниках и монографиях (если таковое вообще есть) до сих пор остается на уровне примеров и «рецептов», за исключением точного теста Фишера (см., например, [32, 33]).

Ниже основное внимание будет уделено именно точным методам анализа дискретных случайных величин, хотя и некоторые приемы асимптотических вычислений также будут приведены.

Технически аппарат, используемый для точного анализа дискретных случайных величин, восходит к работе Р. А. Фишера [28], в которой вероятность попадания в хвост распределения представлялась как отношение двух многочленов, причем числитель был частью знаменателя. Это отношение Фишер использовал для вычисления того, что сейчас называется верхней границей доверительного интервала. Двусторонний вариант такого отношения был предложен в работе [34]. В дальнейшем такое представление использовалось и в других работах, см., например, [35–37]. В работе [38] этот подход был систематизирован и развит с приложениями к различным задачам статистического анализа. В частности, в этой работе вводится понятие доказательной функции (evidence function, или функции  $p$ -значения), которая объединяет три ключевых средства анализа данных:  $p$ -значение, точечную оценку и доверительный интервал. С математической точки зрения это понятие и сопутствующие конструкции довольно элементарны и не требуют знаний, выходящих за пределы вузовского курса алгебры. Однако оно предоставляет эффективную теоретическую и вычислительную среду для точного анализа дискретных данных.

## 2. Основные целочисленные случайные величины

Использование методов теории вероятностей в анализе данных основано на предположении о случайности рассматриваемых исходов. В обсуждаемых ниже вопросах предполагается, что случайность исхода описывается некоторой дискретной случайной величиной. Например, числа в таблицах сопряженности представляют собой дискретные случайные величины, принимающие целочисленные значения. Основные распределения такого типа рассмотрены ниже.

### 2.1. Биномиальное распределение

Допустим, что имеется однородная совокупность, для каждого объекта которой некоторая случайная величина  $y$  принимает значения 0 или 1 с одной и той же вероятностью  $\pi = P(y = 1)$  для каждого объекта (испытания Бернулли;  $\pi$  называется вероятностью успеха). Пусть из этой совокупности извлекается выборка объема  $n$ , и  $y_i$  означает значение случайной величины  $y$  для  $i$ -го объекта, причем будем предполагать, что значения  $y_i$  для разных объектов независимы. Тогда случайная величина  $\mathcal{B} = \sum_i y_i$  есть (случайное) число объектов, для которых  $y = 1$  (биномиальная случайная величина). Плотность распределения этой случайной величины задается равенством

$$P(\mathcal{B} = m; \pi) = C_n^m \pi^m (1 - \pi)^{n-m}, \quad m = 0, 1, 2, \dots, n, \quad C_n^m = \frac{n!}{m!(n-m)!} \quad (1)$$

Нетрудно проверить, что сумма двух независимых биномиальных распределений  $\mathcal{A}$  и  $\mathcal{B}$  с одинаковой вероятностью успеха  $\pi$ , но разным числом испытаний  $n$  и  $m$  соответственно, также имеет биномиальное распределение с вероятностью успеха  $\pi$  и числом испытаний  $n + m$ .

**Пример 1.** Допустим, что для какого-то лекарства исследуется наличие побочных эффектов. Пусть  $\pi$  – неизвестная вероятность того, что у пациентов данной группы будет обнаружено побочное действие лекарства. Будем предполагать, что распределение числа пациентов, у которых будет обнаружен побочный эффект, подчиняется биномиальному распределению  $\mathcal{B}(n, \pi)$ , где  $n$  – число испытуемых.

## 2.2. Полиномиальное распределение

Если в схеме повторных независимых испытаний исходов будет не два, а больше, например,  $m$ , то для описания исхода такого испытания для каждого объекта необходимо вводить  $m$  случайных величин  $y_{ij}$  или одну  $m$ -мерную случайную величину  $y_i = (y_{i1}, y_{i2}, \dots, y_{im})$ , определяемую условиями  $\sum_j y_{ij} = 1$  и  $y_{ij} = 1$ , если для  $i$ -го объекта случился  $j$ -й исход. Очевидно, выполняется равенство  $\sum_{ij} y_{ij} = n$  – полное число проведенных испытаний (или, что тоже самое, число испытуемых объектов). При этом ясно, что независимых случайных величин из  $(y_{i1}, y_{i2}, \dots, y_{im})$  будет  $m - 1$ , так как  $\sum_j y_{ij} = 1$ .

Как и для биномиального распределения, будем считать, что для каждого объекта проводимых испытаний вероятность данного исхода из  $m$  возможных одна и та же  $\pi_j = P(y_{ij} = 1)$ ,  $j = 1, 2, \dots, m$ . Обозначим  $n_j = \sum_i y_{ij}$  – количество объектов, имеющих исход на уровне  $j$ ,  $\mathcal{M} = \sum_i y_i$  – многомерная случайная величина,  $j$ -я компонента которой есть биномиальная случайная величина с вероятностью успеха  $\pi_j$ ,  $\sum_j \pi_j = 1$  (полиномиальное, или мультиномиальное распределение). Тогда плотность полиномиального распределения имеет вид

$$P(\mathcal{M} = (n_1, n_2, \dots, n_m); \pi_1, \pi_2, \dots, \pi_m) = \frac{n!}{n_1! n_2! \dots n_m!} \pi_1^{n_1} \pi_2^{n_2} \dots \pi_m^{n_m}, \quad \sum_i n_i = n \quad (2)$$

**Пример 2.** В условиях примера 1 можно рассматривать более подробно, какие именно побочные явления могут появиться от применения исследуемого лекарства. Например, пусть с вероятностями  $\pi_1, \pi_2, \pi_3$  в данной группе испытуемых может проявиться только один из симптомов: головная боль, жар или мышечный спазм, а с вероятностью  $\pi_4$  не будет никаких побочных явлений,  $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$ . Тогда распределение числа появлений того или иного симптома в группе из  $n$  испытуемых описывается полиномиальным распределением (2).

## 2.3. Отрицательное биномиальное распределение

Пусть  $y_i, i = 1, 2, 3, \dots$  – последовательность независимых испытаний Бернулли с одной и той же вероятностью успеха  $\pi$ ,  $r \geq 1$  – фиксированное целое число. Случайная величина  $\mathcal{NB}$  определяется как число успехов, которые были до наступления  $r$ -й неудачи. Иначе говоря, если  $k + r$  – номер испытания, в котором наступила  $r$ -я неудача, то  $\mathcal{NB} = k$ . Плотность отрицательного биномиального распределения задается равенством

$$P(\mathcal{NB} = k; \pi) = C_{k+r-1}^{r-1} \pi^k (1-\pi)^r, \quad k = 0, 1, 2, \dots \quad (3)$$

Заметим, что отрицательное биномиальное распределение определяется различно и приводит к различным выражениям для плотности вероятности в зависимости от того, означает  $\pi$  вероятность успеха или неудачи и продолжаются испытания до  $r$ -й неудачи или  $r$ -го успеха, а также от того, что подсчитывает случайная величина – число успехов/неудач или число испытаний. Например, если в той же постановке случайная величина  $\mathcal{X}$  подсчитывает число неудач в последовательности испытаний Бернулли с вероятностью успеха  $\pi$ , проводимых до  $r$ -го успеха, то, очевидно, что  $\mathcal{X}$  имеет отрицательное биномиальное распределение и

$$P(\mathcal{X} = k; \pi) = C_{k+r-1}^{r-1} \pi^r (1-\pi)^k, k = 0, 1, 2, \dots$$

Но если  $\mathcal{X}$  подсчитывает число испытаний Бернулли с вероятностью успеха  $\pi$  до наступления  $r$ -го успеха (неудачи), то также получим отрицательное биномиальное распределение (так как  $n = k + r$ )

$$P(\mathcal{X} = n; \pi) = C_{n-1}^{r-1} \pi^r (1-\pi)^{n-r}, n = r, r+1, r+2, \dots$$

$$(P(\mathcal{X} = n; \pi) = C_{n-1}^{r-1} \pi^{n-r} (1-\pi)^r, n = r, r+1, r+2, \dots)$$

Если же  $\mathcal{X}$  подсчитывает число успехов в заданном числе  $n$  испытаний Бернулли, то получаем биномиальное распределение (1).

**Пример 3.** Пусть в условиях примера 1 испытания проводятся последовательно на однородной группе испытуемых до обнаружения побочных эффектов лекарства ровно у  $k$  испытуемых. Число требуемых для этого испытаний имеет отрицательное биномиальное распределение с вероятностью «успеха»  $\pi$  (здесь это вероятность появления побочных эффектов).

## 2.4. Гипергеометрическое распределение

Пусть имеется  $N$  объектов, из которых ровно  $n$  имеют некоторую интересующую нас характеристику (некоторое заболевание, дефект и т. п.). Из этой совокупности извлекается (без возвращения) выборка из  $m$  элементов. Случайная величина  $\mathcal{HG}$  определяется как число объектов, имеющих заданную характеристику в случайной выборке объема  $m$ . Плотность распределения этой случайной величины равна

$$P(\mathcal{HG} = k) = \frac{C_n^k \cdot C_{N-n}^{m-k}}{C_N^m}, k = 0, 1, 2, \dots, m \quad (4)$$

Очевидно, гипергеометрическое распределение можно получить как условное распределение из двух одинаковых независимых биномиальных распределений. А именно, если  $\mathcal{A}$  и  $\mathcal{B}$  независимые биномиальные распределения с одной и той же вероятностью успеха и одним и тем же числом испытаний  $n$ , то распределение одного из них при фиксированной сумме этих случайных величин имеет гипергеометрическое распределение

$$P(\mathcal{A} = k | \mathcal{A} + \mathcal{B} = m) = \frac{C_n^k \cdot C_n^{m-k}}{C_{2n}^m}$$

**Пример 4.** В группе из  $N$  больных некоторым определенным заболеванием имеется  $n$  больных с легким течением и  $N - n$  со средним или тяжелым. Из этой

группы случайным образом выбрано  $m$  больных. Тогда число больных в этой выборке с легким течением заболевания имеет гипергеометрическое распределение.

## 2.5. Распределение Пуассона

Случайная величина  $\mathcal{P}$  имеет распределение Пуассона, если ее плотность задается равенством

$$P(\mathcal{P} = k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots, \quad (5)$$

где параметр  $\lambda > 0$ .

Это распределение используется для представления вероятности наступления редких событий, например, числа ДТП на некотором участке дороги в течение дня, числа новых случаев некоторого заболевания с низкой распространенностью и т. п. Однако фактически его можно рассматривать просто как одно из целочисленных дискретных распределений. Использование этого распределения для описания именно редких событий связано с тем, что можно строго вывести выражение для плотности распределения числа наступивших событий, происходящих в течение некоторого времени или в некоторой пространственной области, предполагая, что

- (1) вероятность появления события в малом интервале времени (или области пространства) пропорциональна размеру этого интервала;
- (2) вероятность двух и более событий в малом интервале времени (области пространства) пренебрежимо мала;
- (3) события в непересекающихся интервалах времени (областях пространства) независимы.

При этих предположениях искомая плотность будет именно плотностью распределения Пуассона. В такой трактовке параметр  $\lambda$  определяет интенсивность наступления событий в течение рассматриваемого промежутка времени (среднее число появлений подсчитываемых событий в течение заданного промежутка времени).

Одно из полезных свойств распределения Пуассона состоит в его аддитивности, т. е. сумма попарно независимых распределений Пуассона с параметрами  $\lambda_i$  также распределена по закону Пуассона с параметром  $\sum_i \lambda_i$ .

**Пример 5.** Пусть в течение суток на станцию скорой помощи поступает в среднем 12 вызовов в связи с ожогами детей. Тогда средняя интенсивность таких обращений равна 0,5 в час. Тогда вероятность поступления  $k$  таких обращений в течение часа вычисляется по формуле (5).

Интересно также, что биномиальное распределение можно получить как условное распределение одного распределения Пуассона при условии постоянства его суммы с другим независимым распределением Пуассона. А именно, если  $\mathcal{P}_1$  и  $\mathcal{P}_2$  – распределения Пуассона с параметрами  $\lambda_1$  и  $\lambda_2$ , то

$$P(\mathcal{P}_1 = k | \mathcal{P}_1 + \mathcal{P}_2 = m) = \frac{m!}{k!(m-k)!} \pi^k (1-\pi)^{m-k}, \quad \pi = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

Важно также, что распределение Пуассона является предельным для отрицательного биномиального при  $r \rightarrow \infty$ . Довольно часто в биологических задачах именно отрицательное биномиальное распределение оказывается более подходящим, чем пуассоновское, см., например, [39–41].

### 3. Степенное представление целочисленной случайной величины

Любую дискретную случайную величину можно считать принимающей значения  $0, 1, 2, \dots$  (до некоторого  $n$  или до бесконечности), делая при необходимости перекодировку ее значений. Большинство таких целочисленных случайных величин можно представить в следующем степенном виде.

Пусть  $\phi$  – положительный параметр,  $f(\phi)$  – сходящийся степенной ряд, в частности, возможно, полином,  $f(\phi) = \sum_{k=0}^{\infty} c(k)\phi^k$ , причем  $c(k) \geq 0$ ,  $k = 0, 1, 2, \dots$ . Обозначим  $\Omega = \{k \mid c(k) > 0\} \neq \emptyset$ . Тогда

$$f(\phi) = \sum_{k \in \Omega} c(k)\phi^k$$

Функция  $f(\phi)$  определяет дискретную (целочисленную) случайную величину  $\mathcal{K}$ , определенную на множестве  $\Omega$  (называемом *носителем* случайной величины  $\mathcal{K}$ , или выборочным пространством для этой случайной величины), следующим равенством ( $\mathcal{K}$  зависит от параметра  $\phi$ )

$$P(\mathcal{K} = k; \phi) = \frac{c(k)\phi^k}{f(\phi)}, k \in \Omega \quad (6)$$

Очевидно, любую целочисленную случайную величину можно представить в виде (6), считая, что  $\phi = 1$ , если она не зависит от какого-либо положительного параметра. Заметим также, что функция  $f(\phi)$  является просто ненормированной производящей функцией случайной величины  $\mathcal{K}$ , которую можно назвать *производящим степенным рядом*, или *производящим полиномом* (если множество  $\Omega$  конечно) для случайной величины  $\mathcal{K}$ . Такое задание случайной величины называется степенным (в англ. литературе *power series distribution* – см. [42, 43]).

Иногда удобно перейти к экспоненциальной форме функции  $f(\phi)$  с параметром  $\beta = \ln(\phi)$

$$h(\beta) = f(e^\beta) = \sum_{k \in \Omega} c(k)e^{\beta k};$$

$$P(\mathcal{K} = k; \beta) = \frac{c(k)e^{\beta k}}{h(\beta)}$$

С помощью функции  $f(\phi)$  (или его экспоненциального варианта  $h(\beta)$ ) можно вычислить числовые характеристики  $\mathcal{K}$  следующим образом.

**Теорема 1.** Если среднее  $\mu$  и дисперсия  $\sigma^2$  случайной величины  $\mathcal{K}$  конечны, то

$$\mu = \phi \frac{d}{d\phi} [\ln f(\phi)] = \frac{f_*(\phi)}{f(\phi)} = \frac{h'(\beta)}{h(\beta)};$$

$$\sigma^2 = \mu + \phi^2 \frac{d^2}{d\phi^2} [\ln f(\phi)] = \frac{f_{**}(\phi)}{f(\phi)} - \left(\frac{f_*(\phi)}{f(\phi)}\right)^2 = \frac{h''(\beta)}{h(\beta)} - \left(\frac{h'(\beta)}{h(\beta)}\right)^2,$$

где  $f_*(\phi) = \sum_{k \in \Omega} kc(k)\phi^k$ ,  $f_{**}(\phi) = \sum_{k \in \Omega} k^2c(k)\phi^k$ .

Доказательство получается прямым вычислением.

Важное свойство порождающих степенных рядов (как и общих производящих функций) состоит в следующем.

**Теорема 2.** Если  $\kappa_1, \kappa_2, \dots, \kappa_N$  – независимые случайные величины, определенные степенными рядами  $f_1(\phi), f_2(\phi), \dots, f_N(\phi)$ , заданными, соответственно, на носителях  $\Omega_1, \Omega_2, \dots, \Omega_N$ , то случайная величина  $\kappa = \kappa_1 + \kappa_2 + \dots + \kappa_N$  определяется степенным рядом

$$f(\phi) = \prod_{k=1}^N f_k(\phi)$$

Доказательство проводится по индукции (см., например, [38, с. 9]).

**Пример 6.** Приведем степенные представления рассмотренных ранее случайных величин.

(1) *Биномиальное распределение*  $\mathcal{B}$  зависит от параметра  $\pi$  (вероятность успеха), относительно которого формула (1) принимает вид

$$P(\mathcal{B} = m; \pi) = C_n^m \pi^m (1 - \pi)^{n-m} = C_n^m \left(\frac{\pi}{1-\pi}\right)^m (1-\pi)^n = C_n^m \left(\frac{\pi}{1-\pi}\right)^m \left(1 + \frac{\pi}{1-\pi}\right)^{-n}$$

Следовательно, для шанса успеха  $\phi = \frac{\pi}{1-\pi}$  это равенство можно представить в виде

$$P(\mathcal{B} = m; \phi) = \frac{C_n^m \phi^m}{(1+\phi)^n}, \quad f(\phi) = (1+\phi)^n$$

(2) *Распределение Пуассона.* Здесь очевидно, что это случайная величина, определяемая степенным рядом, а именно рядом для  $e^\lambda$

$$P(\mathcal{P} = k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda} = \frac{\lambda^k / k!}{e^\lambda}, \quad k = 0, 1, 2, \dots,$$

(3) *Отрицательное биномиальное распределение.* Из равенства  $P(\mathcal{NB} = k; \pi) = C_{k+r-1}^{r-1} \pi^k (1-\pi)^r$ ,  $k = 0, 1, 2, \dots$  сразу получаем

$$P(\mathcal{NB} = k; \phi) = \frac{C_{k+r-1}^{r-1} \phi^k}{f(\phi)}, \quad k = 0, 1, 2, \dots,$$

где  $\phi = \pi$ ,  $f(\phi) = (1-\phi)^{-r} = \sum_{u=0}^{\infty} C_{u+r-1}^{r-1} \phi^u$ .

(4) *Гипергеометрическое распределение.* Для получения степенного представления этой случайной величины удобно перейти к ее обобщению (нецентральному гипергеометрическому распределению) на основе показанной выше связи гипергеометрического и биномиального распределений. А именно, пусть  $\mathcal{A}(n, \pi_1)$  и  $\mathcal{B}(m, \pi_0)$  – взаимно независимые биномиальные распределения. Рассмотрим распределение одного из них при заданной их сумме  $P(\mathcal{A} = k \mid \mathcal{A} + \mathcal{B} = l)$ .

$$\begin{aligned} P(\mathcal{A} = k \mid \mathcal{A} + \mathcal{B} = l) &= \frac{P(\mathcal{A} = k)P(\mathcal{B} = l - k)}{P(\mathcal{A} + \mathcal{B} = l)} = \frac{P(\mathcal{A} = k)P(\mathcal{B} = l - k)}{\sum_{u \in \Omega} P(\mathcal{A} = u)P(\mathcal{B} = l - u)} = \\ &= \frac{\frac{C_n^k \phi_1^k}{(1+\phi_1)^n} \cdot \frac{C_m^{l-k} \phi_0^{l-k}}{(1+\phi_0)^m}}{\sum_{u \in \Omega} \frac{C_n^u \phi_1^u}{(1+\phi_1)^n} \cdot \frac{C_m^{l-u} \phi_0^{l-u}}{(1+\phi_0)^m}} = \frac{C_n^k C_m^{l-k} \left(\frac{\phi_1}{\phi_0}\right)^k \phi_0^l}{\sum_{u \in \Omega} C_n^u C_m^{l-u} \left(\frac{\phi_1}{\phi_0}\right)^u \phi_0^l} = \frac{c(k) \phi^k}{\sum_{u \in \Omega} c(u) \phi^u}, \quad f(\phi) = \sum_{u \in \Omega} c(u) \phi^u, \end{aligned}$$

где  $\Omega = \{l_1, l_1 + 1, \dots, l_2\}$ ,  $l_1 = \max(0, l - m)$ ,  $l_2 = \min(n, l)$ ,  $\phi = \phi_1 / \phi_0 = \frac{\pi_1 / (1 - \pi_1)}{\pi_0 / (1 - \pi_0)}$ ,  $c(k) = C_n^k C_m^{l-k}$ .

Таким образом, нецентральное гипергеометрическое распределение является степенным, а обычное гипергеометрическое распределение получается из этого представления при  $\phi = 1$ . Заметим также, что параметр  $\phi$ , от которого зависит нецентральное гипергеометрическое распределение, представляет собой отношение шансов для параметров распределений  $\mathcal{A}(n, \pi_1)$  и  $\mathcal{B}(m, \pi_0)$ .

(5) *Полиномиальное распределение*. Это вероятностное распределение также является степенным, однако для его представления в этом виде необходимо рассматривать многочлен от многих переменных.

Пусть  $\phi = (\phi_1, \phi_2, \dots, \phi_N)$  – вектор параметров,  $\mathcal{K} = (\mathcal{K}_1, \mathcal{K}_2, \dots, \mathcal{K}_N)$  – вектор из  $N$  дискретных целочисленных случайных величин,  $\mathcal{K}_i = \mathcal{K}_i(\phi_i)$ , принимающий значения из не более чем счетного  $N$ -мерного множества  $\Omega$ . Обозначим (векторное обозначение экспоненты многомерных векторов)

$$\phi^{\mathcal{K}} = \prod_{i=1}^N \phi_i^{\mathcal{K}_i}, \quad \mathbf{k} = (k_1, k_2, \dots, k_N)$$

Важно отметить, что  $\phi^{\mathcal{K}}$  – скаляр, а не вектор. Функция

$$f(\phi) = \sum_{\mathbf{u} \in \Omega} c(\mathbf{u}) \phi^{\mathbf{u}}$$

представляет собой степенной ряд (в частности, полином) от  $N$  переменных. Если  $\mathbf{k}$  –  $N$ -мерный вектор из  $\Omega$ , то  $N$ -мерная случайная величина  $\mathcal{K}$  называется *степенной*, если выполняется равенство

$$P(\mathcal{K} = \mathbf{k}; \phi) = \frac{c(\mathbf{k}) \phi^{\mathbf{k}}}{f(\phi)}$$

Соответствующее экспоненциальное представление имеет вид (штрих означает транспонирование)

$$P(\mathcal{K} = \mathbf{k}; \phi) = \frac{c(\mathbf{k}) \exp(\boldsymbol{\beta} \mathbf{k}')}{h(\boldsymbol{\beta})};$$

$$h(\boldsymbol{\beta}) = \sum_{\mathbf{u} \in \Omega} c(\mathbf{u}) \exp(\boldsymbol{\beta} \mathbf{u}')$$

В частности, для полиномиального распределения (2) получаем его представление в виде степенной случайной величины

$$P(\mathcal{M} = (n_1, n_2, \dots, n_m); \pi_1, \pi_2, \dots, \pi_m) = \frac{n!}{n_1! n_2! \dots n_m!} \pi_1^{n_1} \pi_2^{n_2} \dots \pi_m^{n_m}.$$

В качестве иллюстрации полезности полиномиального представления приведем выражения для вероятности хвостов распределения целочисленной случайной величины. Пусть  $\mathcal{K}$  – целочисленная случайная величина, зависящая от параметра  $\phi$  и определенная степенным рядом  $f(\phi)$  по формуле (6). Тогда, очевидно, имеют место равенства

$$P(\mathcal{K} = k; \phi) = \frac{f(\phi; k)}{f(\phi)}$$

$$P(\mathcal{K} \leq k; \phi) = \frac{f(\phi; \leq k)}{f(\phi)} \quad (7)$$

$$P(\mathcal{K} \geq k; \phi) = \frac{f(\phi; \geq k)}{f(\phi)},$$



где

$$f(\phi; k) = c(k)\phi^k$$

$$f(\phi; \leq k) = \sum_{m \in \Omega, m \leq k} c(m)\phi^m$$

$$f(\phi; \geq k) = \sum_{m \in \Omega, m \geq k} c(m)\phi^m$$

$$f(\phi) = \sum_{m \in \Omega} c(m)\phi^m$$

Таким образом, при известном законе распределения случайной величины  $X$  мы можем точно вычислить вероятность соответствующих левых частей равенств (7), которую можно использовать для нахождения  $p$ -значения.

#### 4. Статистические методы анализа

Методы статистического анализа дискретных данных, в частности таблиц сопряженности  $2 \times 2$ , использующие вероятности попадания в хвост распределения (иначе говоря, вычисление  $p$ -значения), можно разделить на точные и асимптотические методы. Точные методы анализа используют конкретное распределение, которое порождает имеющиеся данные, а асимптотические обращаются к общим критериям, которые являются предельными распределениями для случайных величин, построенных на основе точных распределений. При этом асимптотические критерии не учитывают выборочные схемы получения данных и потому рассматриваются как универсальные. Однако для малых выборок они могут давать довольно заметные ошибки в величине  $p$ -значения или размере доверительного интервала или фактической вероятности попадания в него. Кроме того, существуют методы, занимающие промежуточное положение между точными и асимптотическими, которые основаны на вычислении скорректированных тем или иным способом выражений для  $p$ -значений.

Статистический анализ на основе вычисления вероятности попадания в хвост распределения можно проиллюстрировать на следующем примере. Пусть имеется выборка  $(x_1, x_2, \dots, x_n)$  из нормального распределения с неизвестным средним  $\mu$  и неизвестной дисперсией  $\sigma^2$ . Тогда для проверки гипотезы  $H_0: \mu = \mu_0$  с помощью  $t$ -критерия Стьюдента надо ввести случайную величину ( $\bar{x}$  – выборочное среднее,  $s^2$  – исправленная выборочная дисперсия)

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \quad (8)$$

которая имеет распределение Стьюдента с  $n - 1$  степенью свободы. Вывод о том, согласуется ли гипотеза  $H_0$  с имеющейся выборкой, определяется по тому, насколько велико или мало значение статистики  $t$ . Считая, что рассматривается двусторонний вариант (т. е. альтернативная гипотеза  $H_1: \mu \neq \mu_0$ ), вычисляется значение (8) и его абсолютная величина сравнивается с «критическим значением»  $t_{кр}$  распределения Стьюдента с  $n - 1$  степенями свободы, которое удовлетворяет условию

$$P(|t| > t_{кр}) = \alpha,$$

где  $\alpha$  – заданный уровень значимости (скажем, 0,05). Если вычисленное значение (8) меньше, чем  $t_{кр}$ , то гипотеза  $H_0$  не отвергается, а если больше – отвергается на заданном уровне значимости  $\alpha$ .

Другой подход состоит в том, что по данной выборке вычисляется  $p$ -значение, на основании величины которого делается заключение о принятии/непринятии гипотезы  $H_0$ . А именно, вычисляется вероятность (здесь  $t$  – распределение Стьюдента с  $n - 1$  степенями свободы)

$$p = P(|t| > \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right|). \quad (9)$$

Если эта вероятность мала (например, меньше 0,05), то это является свидетельством против гипотезы  $H_0$ , если же  $p > 0,05$ , то гипотезу  $H_0$  принимают.

Таким образом, гипотеза  $H_0$  принимается или не принимается в зависимости от того, выполняется или не выполняется неравенство  $P(|t| > |t_{\text{набл}}|) > \alpha$ ,  $t_{\text{набл}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ .

Введем индикаторную переменную  $\delta$  скобкой Айверсона  $\delta = [P(|t| > |t_{\text{набл}}|) > \alpha]$  (Iverson's bracket: если  $P$  – некоторое высказывание, то скобка Айверсона от этого высказывания определяется условиями:  $[P] = 1$ , если  $P$  – истинно, и  $[P] = 0$ , если  $P$  ложно). Тогда переменная  $\delta$  определяет нерандомизированный критерий для проверки гипотезы  $H_0$ :  $\delta = [p > \alpha]$ , где  $p$ -значение  $p$  вычисляется по формуле (9). Иначе говоря, гипотеза  $H_0$  принимается тогда и только тогда, когда  $\delta = 1$ . Равносильно можно говорить, что принимается гипотеза  $H_a$ , где  $\delta = [p \leq \alpha] = a$ .

Если для вычисления  $p$ -значения используется некоторая случайная величина, которая (а) строится по выборке; (б) вероятностное распределение которой будет точно известно при выполнении гипотезы  $H_0$ , то ее можно использовать и в качестве обычного статистического критерия для проверки гипотезы  $H_0$ , и в качестве средства для вычисления  $p$ -значения и задания нерандомизированного критерия  $\delta$  с помощью скобки Айверсона. В этом случае  $p$ -значение является вероятностью попадания точно в один (при односторонней альтернативе) или в один из двух хвостов (при двусторонней альтернативе) этой случайной величины. Однако при проверке статистических гипотез для дискретных случайных величин (в частности, для таблиц  $2 \times 2$ ) часто оказывается целесообразным несколько ослабить это правило вычисления  $p$ -значения, что делает необходимым использование  $p$ -значения как самостоятельного средства проверки статистических гипотез.

Степенное представление целочисленной случайной величины позволяет удобно оперировать с ключевыми конструкциями математического аппарата проверки статистических гипотез – уровнем значимости (размером критерия) и мощностью. Кроме того, с проверяемой статистической гипотезой (односторонней или двусторонней) о значении параметра  $\phi$  очевидно связана задача нахождения области, в которой этот параметр находится с заданной вероятностью (доверительная область), которую также можно найти с помощью функции  $p$ -значения. Вначале рассмотрим случай односторонней гипотезы.

#### 4.1. Односторонняя гипотеза.

##### Уровень значимости и мощность критерия

Пусть  $\kappa$  – целочисленная случайная величина, зависящая от параметра  $\phi$ , для которой выдвинута гипотеза  $H_0: \phi \leq \phi_0$  vs  $H_1: \phi > \phi_0$ . Возможность вычисления  $p$ -значения опирается на следующее свойство случайной величины  $\kappa$ .

**Теорема 3.** Если  $\kappa$  – целочисленная случайная величина, зависящая от параметра  $\phi$  и определяемая степенным рядом с носителем  $\Omega$ , и  $k \in \Omega$  отлично от

$\inf \Omega$  и  $\sup \Omega$ , то функция  $P(\kappa \geq k; \phi)$  монотонно возрастает по  $\phi$ , а  $P(\kappa \leq k; \phi)$  – монотонно убывает по  $\phi$ .

Доказательство получается элементарным вычислением производной от каждой вероятности по  $\phi$  с использованием формул (7) (см. подробности в [36, с. 33, 34]).

Таким образом, при фиксированном значении  $k \in \Omega$  правосторонняя область  $\kappa \geq k$  становится тем больше, чем больше значение параметра  $\phi$ . Соответственно, левосторонняя область  $\kappa \leq k$  становится тем больше, чем меньше значение  $\phi$ . Поэтому разумно рассматривать вероятности хвостов  $P(\kappa \geq k; \phi)$  и  $P(\kappa \leq k; \phi)$  как меру справедливости соответствующих односторонних гипотез. А именно, если в результате наблюдения (измерения) над случайной величиной  $\kappa$ , зависящей от параметра  $\phi$ , получено некоторое значение  $k$ , и для некоторого значения  $\phi = \phi_0$  вероятность попадания значений  $\kappa$  в область  $\kappa \geq k$  мала (например, меньше 0,05), можно сделать заключение, что на самом деле значение параметра  $\phi$  должно быть больше, чем  $\phi_0$  (чтобы в соответствии с Теоремой 3 вероятность попадания в область  $\kappa \geq k$  стала больше).

**Определение 1.** Для заданного значения  $\kappa = k$  функция от параметра  $\phi$

$$p(k, \phi) = P(\kappa \geq k; \phi) \quad (10)$$

называется *функцией  $p$ -значения* для односторонних гипотез  $H_0: \phi \leq \phi_0$  vs  $H_1: \phi > \phi_0$ . Соответственно, функция от параметра  $\phi$

$$p(k, \phi) = P(\kappa \leq k; \phi) \quad (11)$$

называется *функцией  $p$ -значения* для односторонних гипотез  $H_0: \phi \geq \phi_0$  vs  $H_1: \phi < \phi_0$ .

**Замечание 1.** Выражения (10), (11) известны также под названиями доказательная функция, доверительная кривая, функция доверительного интервала, конфлюэнтная функция.

**Пример 7.** Пусть в описанных в примере 1 условиях испытание проводилось на 10 пациентах, и у двоих обнаружены побочные явления. Тогда для оценки того, насколько это свидетельствует в пользу наличия побочных эффектов исследуемого лекарства, разумно вычислить вероятность того, что такое или еще большее число испытуемых будет иметь побочные явления от приема этого лекарства. Иначе говоря, необходимо вычислить вероятность  $P(\kappa \geq 2; \phi)$ , которая, однако, остается зависящей от неизвестного параметра  $\phi$ . Рассмотрим зависимость этой вероятности от  $\phi$ . Из равенств (7) следует

$$P(\kappa \geq 2; \phi) = 1 - P(\kappa < 2; \phi) = 1 - (P(\kappa = 0; \phi) + P(\kappa = 1; \phi)) = 1 - \frac{1 + 10\phi}{(1 + \phi)^{10}},$$

т. е. функция  $p$ -значения здесь равна  $p(2, \phi) = 1 - \frac{1 + 10\phi}{(1 + \phi)^{10}}$ . Так как шанс  $\phi$  является монотонно возрастающей функцией от вероятности  $\pi$ , то вместо гипотезы  $H_0: \pi \leq \pi_0$  против конкурирующей гипотезы  $H_1: \pi > \pi_0$  можно рассматривать соответствующую пару гипотез относительно параметра  $\phi$ :  $H_0: \phi \leq \phi_0$  vs  $H_1: \phi > \phi_0$ . Очевидно, в данном примере нас интересует гипотеза малого значения вероятности  $\pi$ . Поэтому, выбрав некоторое пороговое значение  $\phi_0$ , мы должны оценить, каково может быть максимальное значение  $P(\kappa \geq 2; \phi)$  при справедливости гипотезы  $H_0$ , т. е. при выполнении неравенства  $\phi \leq \phi_0$ .

Таким образом, если мы хотим проверить гипотезу  $H_0: \phi \leq \phi_0$  при условии  $\kappa = k$ , то одностороннее точное  $p$ -значение для проверки этой гипотезы можно определить как (индекс  $c$  от *conventional p-value*)

$$p_c(k; \phi_0) = \sup_{\phi \leq \phi_0} P(\kappa \geq k; \phi) = \frac{f(\phi_0; \geq k)}{f(\phi_0)}, \quad (12)$$

т. е.  $p_c(k; \phi_0) = \sup_{\phi \leq \phi_0} p(k; \phi)$ . Аналогично для проверки односторонней гипотезы  $H_0: \phi \geq \phi_0$  при условии  $\kappa = k$  одностороннее точное  $p$ -значение можно принять равным

$$p_c(k; \phi_0) = \sup_{\phi \leq \phi_0} p(k; \phi) = \sup_{\phi \leq \phi_0} P(\kappa \leq k; \phi) = \frac{f(\phi_0; \leq k)}{f(\phi_0)}. \quad (13)$$

Определяемые равенствами (12), (13)  $p$ -значения обладают теми свойствами, которые ожидаются от этих величин: малое  $p$ -значение означает маловероятную справедливость соответствующей гипотезы. Большое  $p$ -значение (обычно больше принятого порогового значения, например, 0,05) свидетельствует о согласовании выборочных данных и выдвинутой гипотезы на заданном уровне порогового значения. Заметим также, что из теоремы 3 следует, что для гипотезы  $H_0: \phi \leq \phi_0$   $p$ -значение (12) является возрастающей функцией параметра  $\phi_0$ , а для гипотезы  $H_0: \phi \geq \phi_0$   $p$ -значение (13) является убывающей функцией этого параметра. Эти свойства выполняются также и для вводимых ниже других способов вычисления  $p$ -значения.

Если мы проверяем гипотезу, что  $H_0: \pi \leq 0,1$ , т. е.  $H_0: \phi \leq 0,111... = 1/9$ , то соответствующее точное  $p$ -значение (12) будет равно

$$P(\kappa \geq 2; 1/9) = 0,264.$$

Следовательно, на основании этого  $p$ -значения можно сделать вывод, что гипотеза  $H_0$  согласуется с имеющимися данными.

Следующее определение вводит другой способ вычисления  $p$ -значения [38, 44, 45].

**Определение 2.** Если  $\kappa$  – дискретная случайная величина, зависящая от параметра  $\phi$ , то *односторонним средним  $p$ -значением (one-sided mid-p-value)* для проверки гипотезы  $H_0: \phi \leq \phi_0$  против альтернативы  $H_1: \phi > \phi_0$  называется функция от  $\phi_0$

$$p_m(k; \phi_0) = p_c(k; \phi_0) - 0,5 \frac{f(\phi_0; k)}{f(\phi_0)} = \frac{f(\phi_0; \geq k) - 0,5f(\phi_0; k)}{f(\phi_0)}, \quad (14)$$

где  $p_c(t; \phi_0)$  – точное  $p$ -значение (12).

Эта величина также обладает теми же свойствами, которые должны быть у  $p$ -значения, в частности, она монотонна по  $\phi$  (см. теорему 3).

Аналогичная формула для левосторонней области очевидна:

$$p_m(k; \phi) = P(\kappa \leq k; \phi) - 0,5P(\kappa = k; \phi); \quad p_m(k; \phi_0) = \frac{f(\phi_0; \leq k) - 0,5f(\phi_0; k)}{f(\phi_0)}. \quad (15)$$

Таким образом, вероятность значения  $\kappa = k$  распределяется поровну между правой и левой областями (хвостами распределения). Очевидно, что односторон-

нее среднее  $p$ -значение (14), (15) всегда меньше, чем точное  $p$ -значение (12), (13) соответственно.

**Пример 7** (продолжение). Одностороннее среднее  $p$ -значение (14) для проверки гипотезы  $H_0: \phi \leq \phi_0$  равно  $0,264 - 0,194/2 = 0,167$ .

Еще один способ введения  $p$ -значения основан на часто применяемой нормальной аппроксимации дискретных распределений. При большом числе проведенных испытаний односторонняя вероятность  $P(\kappa \leq t; \phi)$ , где  $t$  – произвольное число, обычно достаточно хорошо аппроксимируется функцией распределения нормальной (гауссовой) случайной величины с подходящими параметрами. Например, для биномиального распределения  $\kappa$  это следует из интегральной теоремы Муавра – Лапласа [22, 23]

$$P(\kappa \leq t; \pi) = \frac{1}{2} + \Phi\left(\frac{t - n\pi}{\sqrt{n\pi(1-\pi)}}\right),$$

где  $\Phi(x)$  – функция Лапласа, определяемая равенством  $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt$ .

Следовательно, асимптотическое одностороннее  $p$ -значение  $p_a(t; \pi)$  для биномиального распределения можно определить равенствами

$$\begin{aligned} p_a(t; \pi) &= \frac{1}{2} + \Phi(z); \\ p_a(t; \pi) &= \frac{1}{2} - \Phi(z); \end{aligned} \tag{16}$$

для левосторонней области и правосторонней области соответственно, где  $z = \frac{t - n\pi}{\sqrt{n\pi(1-\pi)}}$ . Аналогично будут определяться асимптотические односторонние

$p$ -значения для других дискретных распределений.

**Пример 7** (продолжение). Здесь  $n = 10$ ,  $\pi = 0,1$ ,  $t = 2$ . Значит, асимптотическое правостороннее  $p$ -значение равно  $p_a(2; 0,1) = \frac{1}{2} - \Phi(1,05) = 0,5 - 0,3531 = 0,1469$ .

**Замечание 1.** Выше были введены три варианта односторонних  $p$ -значений: обычный, основанный на вероятности хвоста распределения, среднее  $p$ -значение и асимптотическое  $p$ -значение. Причем последние два, строго говоря, не являются вероятностями хвостов распределения данных. Однако все они обладают теми желательными свойствами, которые позволяют делать выводы о справедливости соответствующей односторонней гипотезы. Для непрерывной случайной величины  $p$ -значение, как правило, является вероятностью некоторого события в изучаемом выборочном пространстве. Для дискретной случайной величины это может быть не так. В целом при использовании того или иного варианта  $p$ -значения для дискретной случайной величины необходимо иметь в виду следующие особенности:

- $p$ -значение есть мера правдоподобия данной гипотезы, принимающая значение от 0 до 1. Оно является случайной величиной на выборочном пространстве;
- $p$ -значение выводится на основе некоторых вероятностных рассуждений, но не обязательно само является вероятностью какого-либо события;
- в частности,  $p$ -значение не всегда равно вероятности того, что нулевая гипотеза ложна и не тождественно уровню значимости или величине ошибки первого рода.

Напомним некоторые известные понятия, представленные в удобном для дальнейшего применения виде. Пусть выдвинута некоторая статистическая гипотеза  $H_0$  о значениях параметра  $\phi$ , от которого зависит случайная величина  $\mathcal{K}$  с носителем  $\Omega$ , и задан уровень значимости  $\alpha_0$ . Нерандомизированный критерий  $\delta$  определяется равенством  $\delta(k; \alpha_0) = [p(k) \leq \alpha_0]$ , где  $p(k)$  –  $p$ -значение, определяемое тем или иным образом.

**Определение 3.** Критической областью  $R(\alpha_0) = R_0$  статистического критерия  $\delta$ , соответствующей уровню значимости  $\alpha_0$ , называется множество значений  $t \in \Omega$ , при которых гипотеза  $H_0$  отвергается:

$$R(\alpha_0) = \{k \in \Omega \mid p(k; \alpha_0) \leq \alpha_0\} = \{k \in \Omega \mid \delta(k; \alpha_0) = 1\}.$$

**Определение 4.** Максимальным достигаемым (фактическим) уровнем значимости критерия  $\delta$ , или *максимальным фактическим размером* критерия  $\delta$ , или *максимальным фактическим уровнем ошибки I-го рода*, называется

$$\alpha = \sup_{H_0} \sum_{k \in \Omega} [p(k) \leq \alpha_0] P(\mathcal{K} = k; \phi), \quad (17)$$

где точная верхняя грань вычисляется по области значений параметра  $\phi$ , описываемой гипотезой  $H_0$ . Например, для гипотезы  $H_0: \phi \leq \phi_0$  имеем

$$\alpha = \sup_{\phi \leq \phi_0} \sum_{k \in \Omega} [p(k) \leq \alpha_0] P(\mathcal{K} = k; \phi).$$

Иначе величину (17) можно описать как максимальную (по  $\phi$ , удовлетворяющему  $H_0$ ) вероятность отвергнуть  $H_0$ , когда она верна.

Важной характеристикой статистических тестов является их устойчивость к заданному уровню значимости, т. е. к заданному размеру критерия.

**Определение 5.** Статистический тест  $\delta$  называется консервативным, если  $\alpha \leq \alpha_0$ , и *либеральным* в противоположном случае.

Иначе говоря, консервативный критерий сохраняет заданный размер (не увеличивает его), а либеральный – нарушает (увеличивает).

**Пример 7** (продолжение). Рассмотрим фактический уровень значимости для каждого рассмотренного выше способа вычисления  $p$ -значения. Примем в качестве номинального уровня значимости обычное значение  $\alpha_0 = 0,05$ . Для рассмотренных выше способов вычисления  $p$ -значения получаем, что уровень значимости вычисляется по формуле ( $u = c, t, a$  соответственно, для обычного  $p$ -значения (12), среднего  $p$ -значения (14) и асимптотического  $p$ -значения (16))

$$\alpha_u(\phi) = \sum_{k \in \Omega} [p_u(k; \phi) \leq \alpha_0] P(\mathcal{K} = k; \phi)$$

и в зависимости от конкретного способа вычисления имеет график, показанный на рис. 1.

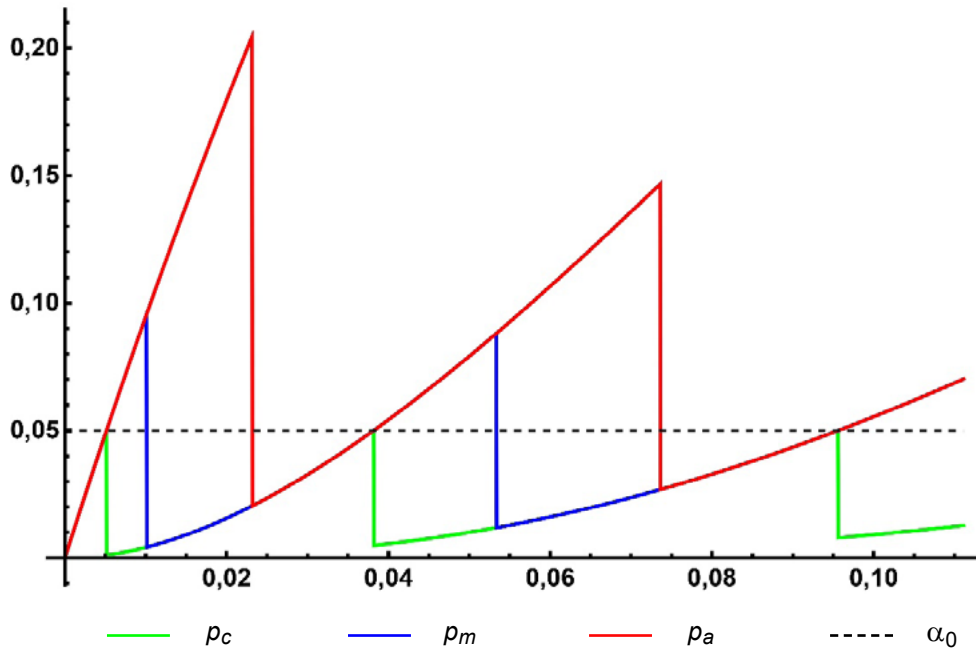


Рис. 1. Уровень значимости критерия, определяемого  $p$ -значением как функция от параметра  $\phi$ . Зеленая линия –  $p$ -значение вычисляется по формуле (12); синяя линия – среднее  $p$ -значение (14); красная линия – асимптотическое  $p$ -значение (16). Части зеленой и синей линий, находящиеся под красной линией, совпадают с ней. Пунктирная линия показывает номинальный выбранный уровень значимости  $\alpha_0 = 0,05$

При этом, как нетрудно проверить, выполняются неравенство  $\sup_{\phi \leq \phi_0} \alpha_c(\phi) = 0,05, \phi_0 = 1/9$ . Следовательно, тест  $\delta$ , основанный на вычислении  $p$ -значения по формуле (12), является консервативным. Напротив, остальные два  $p$ -значения определяют либеральные статистические критерии:  $\sup_{\phi \leq \phi_0} \alpha_m(\phi) = 0,096, \phi_0 = 1/9$  и это значение достигается при  $\phi = 0,01$ ;  $\sup_{\phi \leq \phi_0} \alpha_a(\phi) = 0,2, \phi_0 = 1/9$  достигается при  $\phi = 0,02$ .

Таким образом, вычисление  $p$ -значения по формуле (12) приводит к консервативному статистическому критерию, а по формулам (14) и (16) – к либеральному.

Другая важная характеристика статистического критерия – мощность, т. е. вероятность отклонить нулевую гипотезу, когда она неверна. Имея в виду, что обычно проверяемая гипотеза формулируется как утверждение об отсутствии эффекта (в том или ином смысле – см. [46, 47]), неформально можно сказать, что мощность критерия показывает его способность обнаружить *имеющийся* эффект. В этом контексте уровень значимости показывает возможность критерия (ложно) обнаружить отсутствующий эффект, т. е. сделать вывод из имеющихся данных о наличии эффекта, которого на самом деле нет. Уровень значимости ( $\alpha$ ) и мощность критерия ( $pwr$ ) можно определить равенствами (эти величины определяются для конкретной нулевой гипотезы и ее альтернативы, а также статистического критерия)

$$\alpha = P(\bar{H}_0 \mid H_0), pwr = 1 - P(H_0 \mid H_1)$$

или

$$pwr = P(\text{reject } H_0 \mid H_1 \text{ is true}). \quad (18)$$

Очевидно, что для простых основной и конкурирующей гипотез вычисление мощности не представляет труда, но для сложных гипотез это сделать затруднительно, так как такие гипотезы не определяют однозначно соответствующие вероятностные распределения. Тем не менее в большинстве случаев можно построить функцию двух переменных, которая дает представление о поведении мощности критерия в зависимости от конкретных значений основной и конкурирующей гипотез.

Пусть  $\kappa$  – целочисленная случайная величина с носителем  $\Omega$  и  $H_0: \phi \leq \phi_0$ ,  $H_1: \phi > \phi_0$ . Рассмотрим для каждого  $\phi_1 \leq \phi_0$  критическую область  $R_0(\phi_1) = \{k \in \Omega \mid p(k; \phi_1) \leq \alpha_0\}$ , где  $\alpha_0$  – некоторый назначенный уровень значимости (например, 0,05),  $p(k; \phi)$  –  $p$ -значение, вычисляемое тем или иным образом. Тогда функция мощности (18) принимает вид

$$pwr(\phi_1, \phi_2) = \sum_{k \in R_0(\phi_1)} P(\kappa = k; \phi_2), \quad \phi_2 > \phi_0 \quad (19)$$

Для первоначальной оценки мощности критерия вместо этой функции можно рассматривать функцию  $pwr(\phi_2)$  одной переменной  $\phi_2$ , которая является нижней гранью полной функции (19) по всем значениям  $\phi_1$ , допускаемым гипотезой  $H_0$ , с фиксированным значением  $\phi_2$ . Для такой функции выполняется неравенство  $pwr(\phi_1, \phi_2) \geq \inf_{\phi_1 | H_0} pwr(\phi_1, \phi_2) = pwr(\phi_2)$ , где  $\phi_1 | H_0$  означает, что  $\phi_1$  удовлетворяет гипотезе  $H_0$ . Вместо нахождения этой функции можно воспользоваться тем, что часто существует такое значение  $\phi_1^0 | H_0$ , для которого критическое множество  $R_0(\phi_1^0)$  будет минимальным среди всех  $\phi_1 | H_0$ . Тогда, очевидно, для любого  $\phi_1 | H_0$  выполняется неравенство  $\sum_{k \in R_0(\phi_1)} P(\kappa = k | \phi_2) \geq \sum_{k \in R_0(\phi_1^0)} P(\kappa = k | \phi_2)$ . Например, для рассматриваемого случая гипотез  $H_0: \phi \leq \phi_0$  vs  $H_1: \phi > \phi_0$  таким минимальным множеством будет  $R_0(\phi_0)$  и, следовательно, функция

$$pwr(\phi_2) = \sum_{k \in R_0(\phi_0)} P(\kappa = k | \phi_2) \quad (20)$$

является оценкой снизу для полной функции мощности (19).

**Пример 7** (продолжение). Для обычного  $p$ -значения, вычисляемого по формуле (12) и  $\pi = 0,1$ , т. е.  $\phi_0 = 1/9$ , критическая область равна

$$R_0(\phi_0) = \{4, 5, 6, 7, 8, 9, 10\},$$

а при меньших значениях  $\phi$  к нему добавляются значения 3, 2, 1. Поэтому вычисленная по формуле (20) оценка снизу функции (19) мощности критерия имеет вид

$$pwr(\phi) = \sum_{k=4}^{10} P(\kappa = k; \phi), \quad (21)$$

где значения  $P(\kappa = k; \phi)$  вычисляются по формуле (7). На рис. 2 показан график этой функции на отрезке  $[1/9, 2]$ .



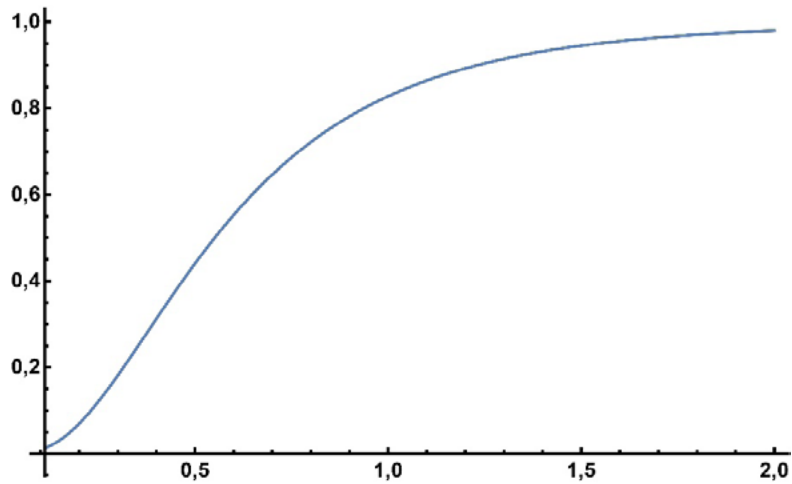


Рис. 2. График функции мощности (21) статистического критерия для примера 1 с консервативным тестом

Разумеется, для конкретных гипотез  $H_0: \phi = \phi_1$  vs  $H_1: \phi = \phi_2$ ,  $\phi_1 < \phi_0 < \phi_2$  функция мощности может значительно отличаться от ее нижней оценки (20). Например, на рис. 3 показаны графики функции мощности  $pwr(\phi_1, \phi)$ ,  $1/9 \leq \phi = \phi_2 \leq 1$ , полученные по формуле (21) при значениях  $\phi_1 = 0,01$  (красная кривая),  $\phi_1 = 0,05$  (синяя),  $\phi_1 = 0,1$  (зеленая кривая). Пунктиром показано значение мощности 0,8.

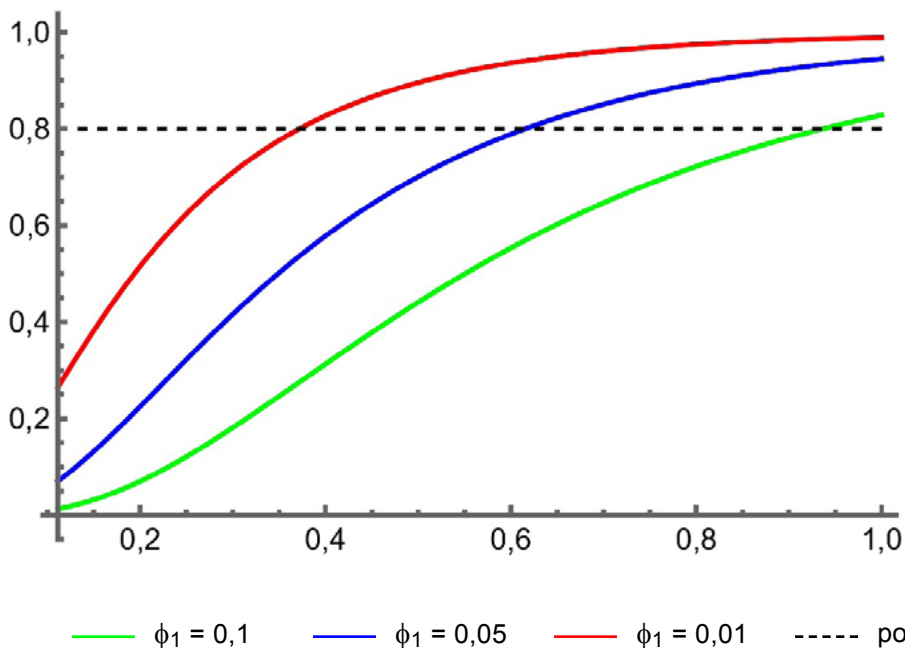


Рис. 3. Сечения  $\phi_1 = \text{const}$  функции мощности (19) для примера 1:  $\phi_1 = 0,01$  (красная кривая);  $\phi_1 = 0,05$  (синяя кривая);  $\phi_1 = 0,1$  (зеленая кривая)

Как видно из рис. 3, чем больше различаются значения  $\phi_1$  и  $\phi_0 = 1/9$ , тем сильнее различаются графики мощности (19) и (20). При  $\phi_1 = 0,1$  функция мощности (19) практически совпадает с функцией (21).

Представление о различии функций мощности (19) для различных статистических тестов на основе вычисления  $p$ -значения по формулам (12), (14), (16) можно увидеть на рис. 4 (для примера 7).

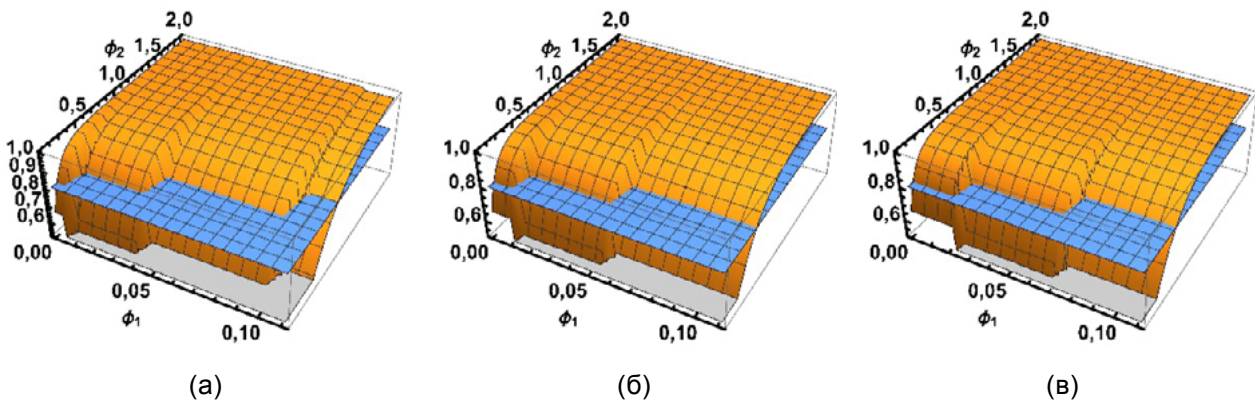


Рис. 4. Поверхность мощности для примера 7 с (а) консервативным тестом; (б) либеральным тестом на основе среднего  $p$ -значения; (в) либеральным тестом на основе асимптотического  $p$ -значения. Голубым цветом показан уровень мощности 0,8

Из рис. 1 и 4 видно, что статистический критерий на основе среднего  $p$ -значения занимает промежуточное положение между точным и асимптотическим критериями не только по величине достигаемого уровня значимости, но и по мощности. В этом смысле его можно считать приемлемым компромиссом в тех случаях, когда точный критерий на основе (12) слишком консервативен, а на основе (16) – слишком либерален.

В целом можно отметить, что вычисление фактического размера критерия и его мощности становится все более распространенным в медико-биологической практике. Разработка новых алгоритмов и теоретических методов такого вычисления делает эти задачи значительно более доступными даже для пользователей, не имеющих глубоких знаний в математической статистике. При этом, несмотря на наличие общей теории и доступность программных средств вычисления характеристик статистических критериев, корректное применение этих алгоритмов, методов и программ предполагает точное понимание всех необходимых шагов таких вычислений и учет специфики конкретной ситуации.

#### 4.2. Односторонние доверительные области

Важной частью анализа целочисленных случайных величин и, в частности, таблиц сопряженности  $2 \times 2$ , является нахождение доверительных интервалов для неизвестных параметров этих распределений. Проверки гипотезы  $H_0: \phi = \phi_0$  о точном значении такого параметра или гипотез  $H_0: \phi \leq \phi_0$ ,  $H_0: \phi \geq \phi_0$  о его односторонней границе часто недостаточно для оценки истинного значения этого параметра. Например, пусть сравнивается эффективность нового метода лечения некоторого заболевания и  $\omega$  – шанс выздоровления после применения этого метода лечения. Проверяемая гипотеза может предполагать отсутствие эффективности этого метода, т. е.  $H_0: \phi \leq 1,0$  против гипотезы  $H_1: \phi > 1,0$ , где  $\phi$  – шанс выздоровления. Предполагая биномиальную модель описания выздоровления, мы получим, что при  $n = 100$  испытуемых и  $k = 55$  выздоровевших  $p$ -значение (12) равно 0,184, а при  $n = 1\,000$  испытуемых и  $k = 550$  выздоровевших это же  $p$ -значение равно 0,00009. При этом в обоих случаях частота выздоровления одна и та же 0,55. Таким образом, в зависимости от объема исследований мы можем сделать противоположные выводы: для  $n = 100$  испытуемых и  $k = 55$  выздоровевших можно считать, что лечение неэффективно, а при  $n = 1\,000$  испытуемых и  $k = 550$  выздоровевших, что оно значимо эффективно.

Понятно, что эти противоположные выводы возникают вследствие значительного различия числа испытуемых в двух опытах. Однако эти вычисления  $p$ -значения не дают возможности оценить, *насколько* эффективно лечение. Как таковое  $p$ -значение показывает, насколько вероятно наблюдаемое значение отклика в предположении выполнения гипотезы  $H_0$ , но никак не описывает насколько велик и *предметно значим* сам эффект, приводящий к такому значению отклика. Иначе говоря, кроме величины  $p$ -значения, необходимо иметь и оценку величины интересующего исследователя эффекта, например, вероятности выздоровления из примера выше. Одним из таких методов является построение доверительных интервалов для интересующей нас величины.

Для односторонних гипотез соответствующие доверительные интервалы будут ограничены только с одной стороны, т. е. будут лучами на числовой прямой. Рассмотренный ниже метод построения таких интервалов представляет собой обращение критерия проверки гипотезы  $H_0$  [46, 47] и удобен для рассмотрения именно дискретных случайных величин.

Пусть рассматривается гипотеза  $H_0: \phi \leq \phi_0$  ( $H_0: \phi \geq \phi_0$ ), где  $\phi$  – параметр, от которого зависит целочисленная случайная величина  $\kappa$ . В результате наблюдения получено значение  $\kappa = k$ . Используя эту информацию, требуется найти такую область изменения параметра  $\phi$ , в которой с заданной вероятностью  $(1 - \alpha_0)$ ,  $0 < \alpha_0 < 1$ , найдется значение параметра  $\phi_0$ .

**Определение 6.** *Правосторонней (левосторонней) доверительной областью уровня  $(1 - \alpha_0)$ ,  $0 \leq \alpha_0 \leq 1$ , для параметра  $\phi_0$  называется множество*

$$\{\phi_0 \mid H_0: \phi \leq \phi_0 \text{ не отвергается на уровне значимости } \alpha_0\}, \quad (22)$$

соответственно

$$\{\phi_0 \mid H_0: \phi \geq \phi_0 \text{ не отвергается на уровне значимости } \alpha_0\}. \quad (23)$$

Поскольку во всех рассматриваемых определениях  $p$ -значения они будут монотонны по параметру  $\phi$  (для разных гипотез монотонность будет разной – см. теорему 3), то множество (22) ограничено снизу, а множество (23) – ограничено сверху. Минимальное (максимальное) значение множества (22) (соответственно, (23)) называется *левой*, или *нижней (правой, или верхней)* доверительной границей параметра  $\phi_0$ .

Так как принятие или непринятие гипотезы  $H_0$  производится на основе величины  $p$ -значения, то такие доверительные области будут различными в зависимости от способа вычисления  $p$ -значения. Например, для обычного выражения (12) для  $p$ -значения гипотеза  $H_0: \phi \leq \phi_0$  будет приниматься при выполнении неравенства

$$\frac{f(\phi_0; \geq k)}{f(\phi_0)} \geq \alpha_0.$$

При этом так как левая часть неравенства является монотонно возрастающей непрерывной функцией от  $\phi_0$ , изменяющейся от 0 до 1, то существует значение  $\phi_0$ , для которого выполняется равенство  $f(\phi_0; \geq k) = \alpha_0 f(\phi_0)$ , которое и является левой границей доверительной области (22). Таким образом, левая (правая) граница доверительной области (22) (соответственно, (23)) уровня  $(1 - \alpha_0)$ ,  $0 < \alpha_0 < 1$ , является решением уравнения

$$\sum_{m \geq k} c(m) \phi^m = \alpha_0 \sum_{m \in \Omega} c(m) \phi^m \quad (24)$$

соответственно,

$$\sum_{m \leq k} c(m) \phi^m = \alpha_0 \sum_{m \in \Omega} c(m) \phi^m. \quad (25)$$

**Замечание 3.** Множества (22), (23) можно представить в виде  $\{\phi_0 \mid p(k; \phi_0) \geq \alpha_0\}$ , где  $p(k; \phi_0)$  – соответствующее  $p$ -значение. Следовательно, доверительная область есть множество  $p^{-1}(k; [\alpha_0; 1])$  (прообраз отрезка по второй переменной). Если рассматривать обычное определение  $p$ -значения (12), (13), то по теореме 3 функция  $p(k; \phi_0)$  монотонна по  $\phi_0$ , что гарантирует, что доверительная область связна, т. е. представляет собой интервал (бесконечный с одной стороны).

Геометрически граница доверительной области получается пересечением графика функции  $p(k; \phi)$  и горизонтальной линии  $\alpha_0$ .

**Пример 7** (продолжение). Для гипотезы  $H_0: \phi \leq \phi_0$ ,  $\alpha_0 = 0,05$  получаем 95 %-ю доверительную область для параметра  $\phi_0$ , решая уравнение (24) или находя точку пересечения графика функции  $p(2; \phi)$ , определяемой равенством (12), с линией  $\alpha_0 = 0,05$ . Аналогично для гипотезы  $H_0: \phi \geq \phi_0$ , 95 %-я доверительная область для параметра  $\phi_0$  находится решением уравнения (25) или поиском точки пересечения графика функции  $p(2; \phi)$ , определяемой равенством (13), с линией  $\alpha_0 = 0,05$  (см. рис. 5).

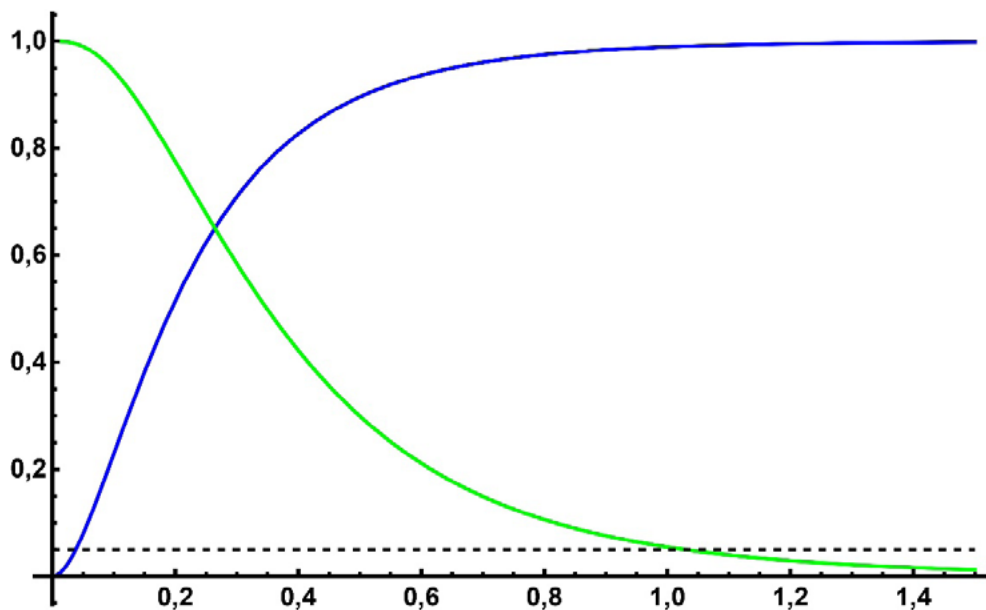


Рис. 5. График  $p(2; \phi)$  для примера 7 и его пересечение с прямой  $\alpha_0 = 0,05$ . Синяя линия –  $p$ -значение для гипотезы  $H_0: \phi \leq \phi_0$ ; зеленая – для гипотезы  $H_0: \phi \geq \phi_0$

Искомое значение для левой границы доверительной области равно 0,038. Следовательно, доверительная область есть луч  $[0,038; \infty)$ . Для правой границы 1,028, доверительная область есть отрезок  $[0; 1,028]$ . Заметим также, что в отрезке  $[0,038; 1,028]$  параметр  $\phi_0$  находится с вероятностью  $0,95^2 = 0,9025$ .

Аналогично находятся границы доверительных областей для других способов вычисления  $p$ -значения. А именно, для среднего  $p$ -значения (14), (15) границы доверительной области уровня  $(1 - \alpha_0)$ ,  $0 < \alpha_0 < 1$ , находятся как решения уравнений

$$\begin{aligned} f(\phi; \geq k) - 0,5f(\phi; k) &= \alpha_0 f(\phi), \\ f(\phi; \leq k) - 0,5f(\phi; k) &= \alpha_0 f(\phi). \end{aligned} \quad (26)$$

Для асимптотического вычисления  $p$ -значения (16) односторонние границы доверительных областей для параметра  $\pi$  находятся как решения уравнения (меньший корень дает левую границу, больший – правую)

$$(t - n\pi)^2 = z_{1-\alpha_0}^2 n\pi(1 - \pi),$$

где  $z_{1-\alpha_0}$  – квантиль уровня  $1 - \alpha_0$  стандартного нормального распределения. Явное решение этого уравнения имеет вид

$$\pi = 0,5(b \pm \sqrt{b^2 - 4ac}), \quad b = \frac{2t + z_{1-\alpha_0}^2}{n + z_{1-\alpha_0}^2}, \quad c = \frac{t^2}{n(n + z_{1-\alpha_0}^2)} \quad (27)$$

**Пример 7** (продолжение). Для одностороннего среднего  $p$ -значения левая и правая границы доверительных областей, соответственно, равны 0,053 и 0,876. Для асимптотического  $p$ -значения эти границы таковы (в отличие от предыдущих случаев здесь это границы для биномиальной вероятности  $\pi$ ): 0,086 и 0,399. Для шанса  $\phi$  эти значения равны 0,094 и 0,664 соответственно.

При сопоставлении односторонних областей для различных способов вычисления  $p$ -значения необходимо учитывать, что среднее и асимптотическое  $p$ -значения либеральны и, следовательно, реальная вероятность того, что оцениваемый параметр находится в определяемой ими области будет меньше заданного уровня доверительной вероятности. Иначе говоря, они дают более оптимистичную доверительную область, чем консервативное  $p$ -значение (12), (13).

**Замечание 4.** Как мы видели, для определения фактического  $p$ -значения для данного наблюдаемого значения  $\kappa = k$  и для определения доверительной области параметра распределения  $\kappa$  используется одна и та же величина – вычисляемое тем или иным образом  $p$ -значение. График зависимости  $p$ -значения от параметра  $\phi$  позволяет найти как величину  $p$ -значения для проверяемой гипотезы, так и доверительную область для этого параметра. Например, пусть рассматривается гипотеза  $H_0: \phi \leq \phi_0$  с уровнем значимости  $\alpha_0 = 0,05$  и наблюдаемым значением  $\kappa = k$  случайной величины  $\kappa$ , зависящей от параметра  $\phi$ . Построим график  $p(k; \phi)$  как функции от  $\phi$ . Тогда  $p$ -значение для проверки гипотезы  $H_0$  находится как ордината точки пересечения этого графика с прямой  $\phi = \phi_0$ , а левая граница односторонней доверительной области – как абсцисса точки пересечения этого графика с прямой  $\alpha = \alpha_0$  (см. рис. 6).

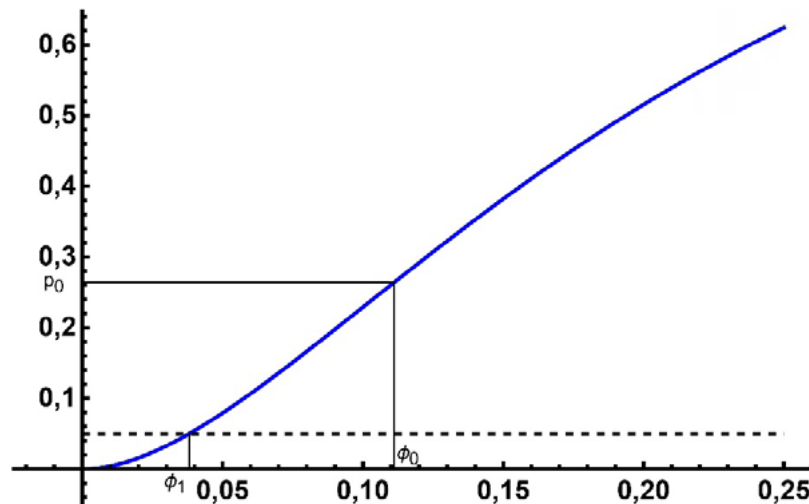


Рис. 6. График  $p$ -значения для проверки гипотезы  $H_0: \phi \leq \phi_0$  при заданном значении  $\kappa = k$  случайной величины  $\kappa$  (синяя линия), а также нахождение соответствующего уровня значимости  $p_0$  и правосторонней доверительной области с левой границей  $\phi_1$ . Пунктирная линия отмечает уровень значимости  $\alpha_0$

### 4.3. Двусторонняя гипотеза

При двустороннем оценивании рассматривается гипотеза  $H_0: \phi = \phi_0$  с различными альтернативами, обычно  $H_1: \phi \neq \phi_0$ . В практическом анализе данных такие гипотезы более естественны, так как не предполагают какую-либо априорную информацию о направленности эффекта. Например, при сравнении двух методов лечения исходная нулевая гипотеза состоит в том, что оба метода имеют одинаковую эффективность. Альтернатива может быть как отрицанием этого предположения, так и выражением большей эффективности какого-то из этих методов.

На основании результатов для одностороннего оценивания можно ожидать, что для проверки справедливости гипотезы  $H_0$  необходимо найти такую функцию  $p(k, \phi)$  вычисления  $p$ -значения, для которой будут выполняться следующие условия:

- $p(k, \phi)$  непрерывная функция от  $\phi$ , принимающая значения от 0 до 1;
- $p(k, \phi)$  возрастает от 0 при изменении  $\phi$  от 0 до некоторого значения  $\phi_1$  и убывает до 0, начиная с некоторого значения  $\phi_2$  (возможно, совпадающего с  $\phi_1$ ).

Тогда как малые, так и большие значения  $\phi$  будут иметь малые значения  $p(k, \phi)$ , что можно считать свидетельством против гипотезы  $H_0: \phi = \phi_0$ .

В качестве такой функции можно выбрать функцию удвоения вероятности минимального хвоста для данного значения  $\kappa = k$  (twice the smaller tail, или double tail method, см., например, [35, 38]), полагая ее равной 1, если удвоенная минимальная вероятность хвоста будет больше 1. Таким образом, положим

$$p(k; \phi) = \min\{2 \min\{P(\kappa \leq k; \phi), P(\kappa \geq k; \phi)\}, 1\}. \quad (28)$$

На рис. 7 показан график этой функции для примера 7.

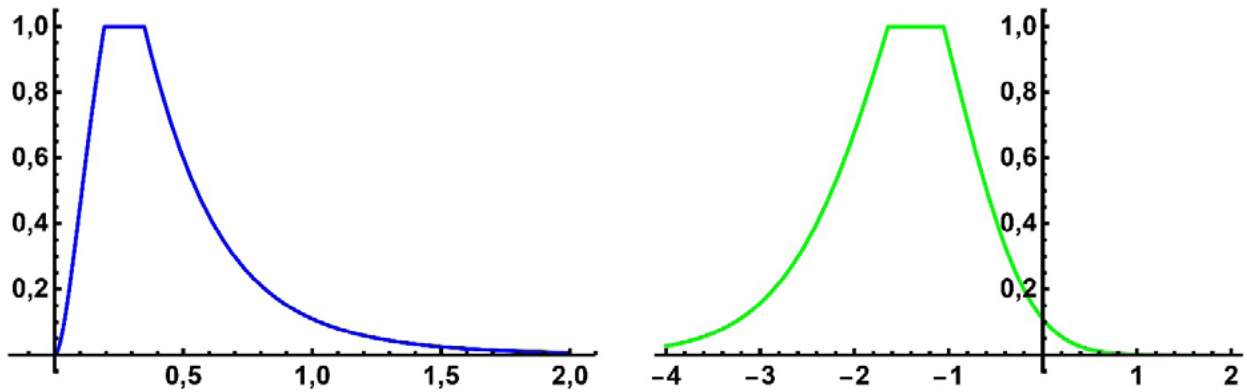


Рис. 7. График функции  $p(2, \phi)$  для примера 7 относительно  $\phi$  (слева) и относительно переменной  $\beta = \ln \phi$  (справа)

С помощью функции (28) можно следующим образом решить обычные задачи оценивания в их двустороннем варианте [38]:

- Для гипотезы  $H_0: \phi = \phi_0$  при  $\kappa = k$  двустороннее  $p$ -значение определяется как  $p(k, \phi_0)$ .
- Двусторонний доверительный интервал уровня  $(1 - \alpha_0)$ ,  $0 \leq \alpha_0 \leq 1$ , для параметра  $\phi_0$  есть множество таких значений  $\phi$ , для которых  $H_0: \phi = \phi_0$  не отвергается на уровне  $\alpha_0$ , т. е. это множество  $\{\phi \mid p(k; \phi) \geq \alpha_0\}$ , или  $p^{-1}(k; [\alpha_0; 1])$ , как и для одностороннего случая. Границы этого интервала находятся как решение уравнения  $p(k; \phi) = \alpha_0$ .
- Значение  $\hat{\phi}(k)$  которое содержится во всех двусторонних доверительных интервалах при  $0 \leq \alpha_0 \leq 1$ , будем называть *точечной оценкой* параметра  $\phi$ . Как следует из рис. 7, в этом подходе такое число может быть неединственным.

**Пример 7** (продолжение). Вычислим двустороннее  $p$ -значение для гипотезы  $H_0: \phi = 1/9$  при  $\kappa = 2$ :

$$p(10; 1/9) = \min\{2\min\{P(\kappa \leq 2; 1/9), P(\kappa \geq 2; 1/9)\}, 1\} = 0,528.$$

Двусторонний доверительный интервал для  $\phi$  с доверительной вероятностью 0,95 для  $p$ -значения, определенного равенством (28), есть отрезок  $[0,026; 1,253]$ . Вследствие наличия постоянного участка для функции (28) точечная оценка, описанная выше, определена неоднозначно – пересечение всех доверительных интервалов есть отрезок  $[0,194; 0,349]$ .

Из свойств одностороннего  $p$ -значения и формулы (28) следует, что выполняются следующие полезные свойства:

- Если  $1 > \alpha_{01} > \alpha_{02} > 0$ , то двусторонний тест, значимый на уровне  $\alpha_{02}$ , будет также значим и на уровне  $\alpha_{01}$ .
- Вложенность доверительных областей: если  $1 > \alpha_{01} > \alpha_{02} > 0$ , то двусторонняя доверительная область уровня  $(1 - \alpha_{01})$  содержится в двусторонней доверительной области уровня  $(1 - \alpha_{02})$ .
- Связность доверительной области: для любого  $0 < \alpha_0 < 1$  двусторонняя доверительная область уровня  $(1 - \alpha_0)$  представляет собой интервал.

Имея в виду введенные выше выражения для одностороннего среднего (14)–(15) и асимптотического (16)  $p$ -значений, можно определить аналогичные двусторонние варианты. Например, двусторонний вариант среднего  $p$ -значения определяется равенством



$$p_m(k; \phi_0) = 2\min\{P(\kappa < k; \phi_0), P(\kappa > k; \phi_0)\} + P(\kappa = k; \phi_0), \quad (29)$$

так как в силу равенства  $P(\kappa < k; \phi) + P(\kappa > k; \phi) + P(\kappa = k; \phi) = 1$  правая часть равенства (29) не будет превышать 1.

Асимптотическое двустороннее  $p$ -значение можно определить равенством

$$p_a(k; \pi) = 2\min\left\{\frac{1}{2} - \Phi(z), \frac{1}{2} + \Phi(z)\right\}, \quad z = \frac{k - n\pi}{\sqrt{n\pi(1-\pi)}}. \quad (30)$$

**Пример 7** (продолжение). Построим графики  $p$ -значений и найдем 95 %-е доверительные интервалы для формул (29), (30) для проверки гипотезы  $H_0: \phi = 1/9$  при  $\kappa = 2$ . Получаем (рис. 8):

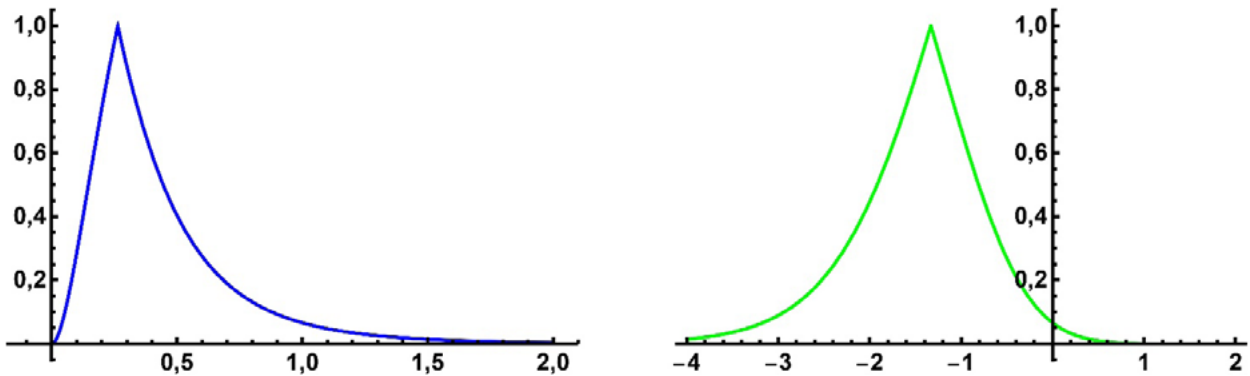


Рис. 8. График функции  $p_m(2, \phi)$  для примера 7 в исходном масштабе (слева) и в логарифмическом относительно переменной  $\beta = \ln \phi$  (справа)

95 %-й доверительный интервал для значения  $\phi$  относительно  $p$ -значения (29) равен  $[0,036; 1,081]$ . Точечная оценка этого параметра равна 0,264 и достигается в точке максимума функции  $p_m(2, \phi)$  (рис. 9).

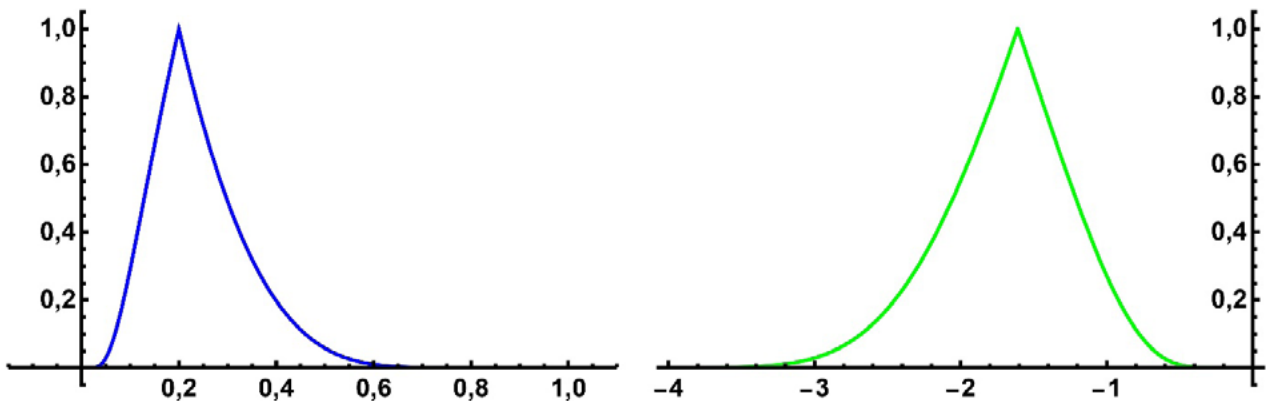


Рис. 9. График функции  $p_a(2, \pi)$  для примера 7 в исходном масштабе (слева) и в логарифмическом относительно переменной  $\beta = \ln \pi$  (справа)

95 %-й доверительный интервал для значения  $\pi$  относительно  $p$ -значения (30) равен  $[0,057; 0,510]$ . Точечная оценка этого параметра равна 0,2 и достигается в точке максимума функции  $p_a(2, \phi)$  (рис. 9).

Можно отметить, что для двустороннего среднего  $p$ -значения (29) точка  $\hat{\phi}$  максимума функции  $p_m(k, \phi)$  удовлетворяет равенству  $f(\hat{\phi}; > k) = f(\hat{\phi}; < k)$  и называется



несмещенной медианной оценкой параметра  $\phi$ . Она, очевидно, содержится в каждом доверительном интервале, т. е. является, таким образом, точечной оценкой параметра  $\phi$ . Для рассматриваемого примера 7 несмещенная медианная оценка параметра  $\pi$  равна 0,264.

#### 4.4. Вероятность покрытия и средняя длина доверительного интервала

Как мы видели, ключевые характеристики статистических выводов –  $p$ -значение и доверительный интервал можно получить из одной функции, а именно, функции  $p$ -значения, вычисление которой может быть определено различными способами, однако в большинстве случаев используется вероятность попадания в хвосты некоторого распределения. Одна из важных характеристик, которую можно использовать при теоретическом анализе того или иного метода вычисления функции  $p$ -значения, называется *вероятность покрытия*. Она позволяет определить, насколько хорошо данная функция  $p$ -значения задает доверительные интервалы для интересующего исследователя параметра. Смысл понятия *вероятность покрытия доверительным интервалом* состоит в том, что это есть доля тех доверительных интервалов (для заданного параметра  $\phi$ ) среди всех доверительных интервалов, которые для фиксированного значения параметра  $\phi$  содержат этот параметр.

Более формально определение можно дать следующим образом. Пусть некоторый параметр  $\phi$  принимает значения из множества  $\Phi$  (например, для биномиальной вероятности  $\Phi = (0; 1)$ ) и его истинное значение равно  $\phi_0$ , т. е.  $\phi_0 \in \Phi$ . После анализа экспериментальных данных делается предположение о том, что на самом деле параметр содержится в более узком множестве  $CS$  (confidence set)  $\phi_0 \in CS \subset \Phi$ . Это предположение может быть верным или неверным, и общепринятым подходом является задание некоторого уровня вероятности, с которым это утверждение будет истинно (*доверительная вероятность*). Как правило, доверительное множество  $CS$  выбирается в соответствии с некоторым правилом (процедурой, формулой)  $\xi$ , которое зависит от экспериментальных данных и, следовательно, имеет случайный характер, который управляется неизвестной функцией распределения  $F$ , порождающей эти данные. В свою очередь, функция  $F$  зависит от параметра  $\phi$ .

**Определение 7.** Вероятность покрытия доверительными множествами  $CS$ , выбираемыми в соответствии с правилом  $\xi$ , истинного значения  $\phi_0$  параметра  $\phi$ , называется вероятностью  $C(CS, \phi_0)$ , определяемой равенством

$$C(\phi_0) = P_{\phi_0}(\phi_0 \in CS(\xi)), \quad (31)$$

где множества  $CS(\xi)$  берутся по всем допустим значениям правила  $\xi$ , а  $P_{\phi_0}$  означает, что вероятность вычисляется с помощью функции распределения  $F$  при значении параметра  $\phi = \phi_0$ . Эта вероятность называется также фактическим уровнем (вероятностью) покрытия параметра  $\phi_0$ .

**Пример 8.** Если  $CI(\phi_0, \alpha_0)$  – доверительный интервал уровня  $1 - \alpha_0$  для параметра  $\phi_0$  случайной величины  $\mathcal{T}$  с носителем  $\Omega$ , то вероятность покрытия (31) можно представить в виде

$$C(\phi_0) = \sum_{\omega \in \Omega} [\phi_0 \in CI(\phi_0, \alpha_0)] P_{\phi_0}(\mathcal{T} = \omega), \quad (32)$$

где  $[ ]$ , как и раньше, скобка Айверсона.

Например, если рассматривается задача оценки биномиальной вероятности  $\pi$  в  $n$  испытаниях, то формула (32) принимает вид

$$C(\pi_0) = \sum_{k=0}^n [\pi_0 \in CI(\pi_0, \alpha_0, k)] C_n^k \pi_0^k (1-\pi_0)^{n-k},$$

где  $CI(\pi_0, \alpha_0, k)$  – доверительный интервал для биномиальной вероятности при условии, что наблюдается  $k$  успехов в  $n$  испытаниях.

Таким образом, величина вероятности покрытия существенно зависит от того, каким образом будет построен доверительный интервал и, следовательно, от того, какой выбрана функция  $p$ -значения.

**Пример 7** (продолжение). Ниже показаны графики вероятности покрытия доверительными интервалами, построенными на основе обычной, средней и асимптотической двусторонней функций  $p$ -значения (соответственно (28), (29), (30)). Особенность дискретных распределений в том, что такие графики имеют пилообразный характер. При этом существенно, что при заданном уровне значимости вероятность покрытия может быть не меньше заданного доверительного уровня при любых значениях оцениваемого параметра (в заданном множестве) или иногда меньше этого уровня. В первом случае говорят о консервативном методе задания доверительного интервала, во втором о либеральном. Как видно из рис. 10, ситуация по отношению к вероятности покрытия такая же, как и ранее для статистического теста на основе соответствующей функции  $p$ -значения: доверительные интервалы, построенные по обычному  $p$ -значению (12), – консервативны, по среднему и асимптотическому  $p$ -значению – либеральны, причем доверительный интервал по среднему  $p$ -значению занимает промежуточное положение между иногда чрезмерно консервативным доверительным интервалом по обычному  $p$ -значению и чрезмерно либеральным по асимптотическому  $p$ -значению.

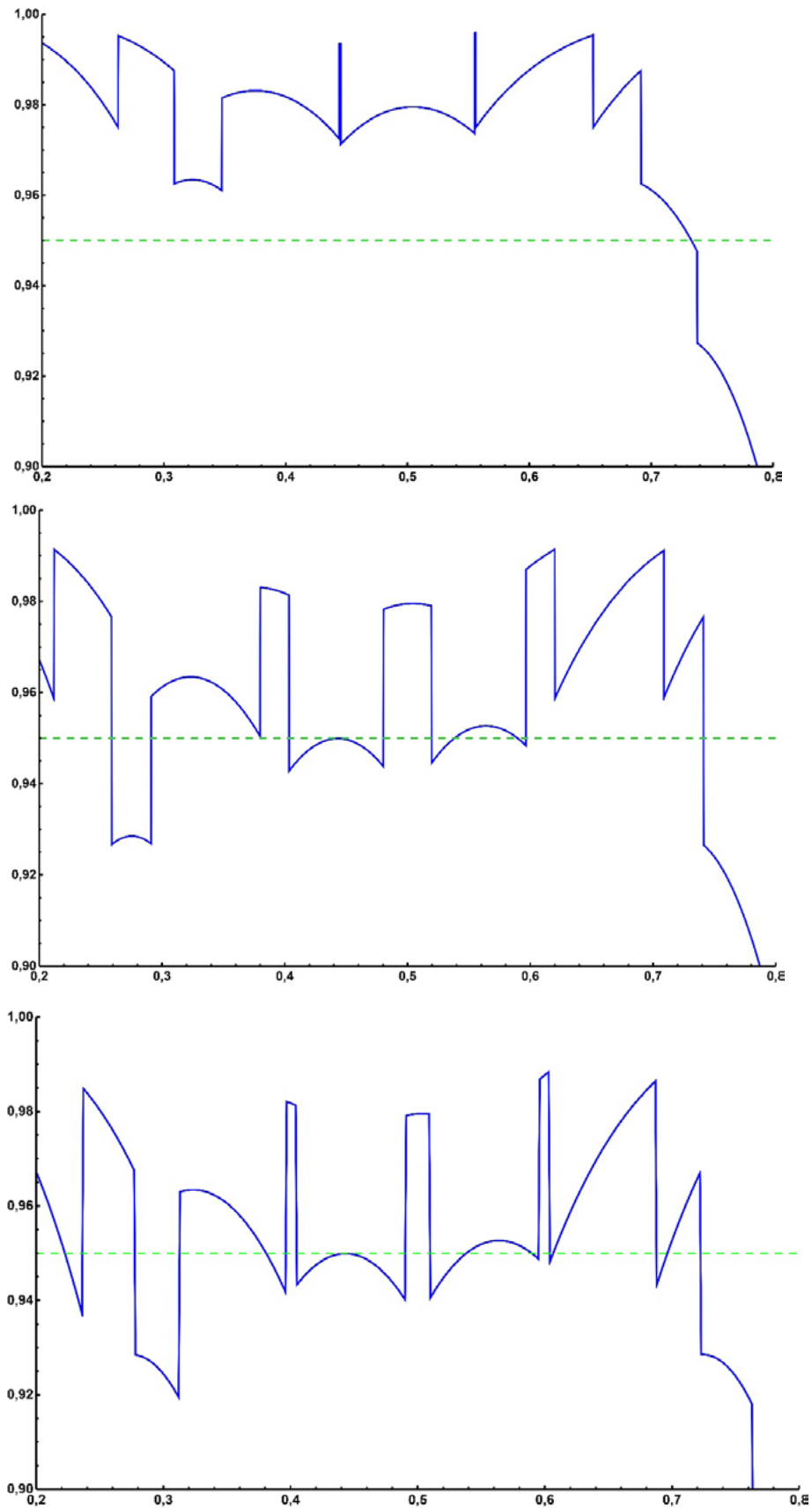


Рис. 10. Графики вероятности покрытия доверительным интервалом неизвестного значения параметра биномиального распределения для примера 7. Ось абсцисс – возможные значения вероятности оцениваемого параметра, ось ординат – соответствующая вероятность покрытия. Сверху вниз для доверительных интервалов на основе обычной функции  $p$ -значения (28), функции среднего  $p$ -значения (29), функции асимптотического  $p$ -значения (30)

С понятием «вероятность покрытия» тесно связана величина средней длины доверительного интервала, который строится по заданному правилу.

**Определение 8.** Средней длиной доверительного интервала для фиксированного значения оцениваемого параметра  $\phi = \phi_0$ , определяемого данным правилом, называется среднее значение длин этих интервалов, содержащих значение  $\phi_0$ .

**Пример 8** (продолжение). Из определения 8 следует, что в обозначениях примера 8 средняя длина  $\text{len}(\phi_0)$  доверительных интервалов, содержащих параметр  $\phi_0$ , вычисляется по формуле

$$\text{len}(\phi_0) = \sum_{\mu \in \Omega} [\phi_0 \in CI(\phi_0, \alpha_0)] \|CI(\phi_0, \alpha_0)\| P_{\phi_0}(T = \phi_0), \quad (33)$$

где  $\|CI(\phi_0, \alpha_0)\|$  – длина интервала  $CI(\phi_0, \alpha_0)$ . В частности, для примера 7 оценки биномиальной вероятности эта формула принимает вид

$$\text{len}(\pi_0) = \sum_{k=0}^n [\pi_0 \in CI(\pi_0, \alpha_0, k)] \|CI(\pi_0, \alpha_0, k)\| C_n^k \pi_0^k (1-\pi_0)^{n-k}$$

Понятно, что вероятность покрытия доверительными интервалами и средняя длина доверительного интервала могут служить ориентирами в выборе подходящего способа вычисления доверительного интервала, однако на практике предпочтение отдается вероятности покрытия. На рис. 11 показаны графики средней длины доверительных интервалов для трех способов определения функции  $p$ -значения (28), (29) и (30) при уровне значимости  $\alpha_0 = 0,05$ .

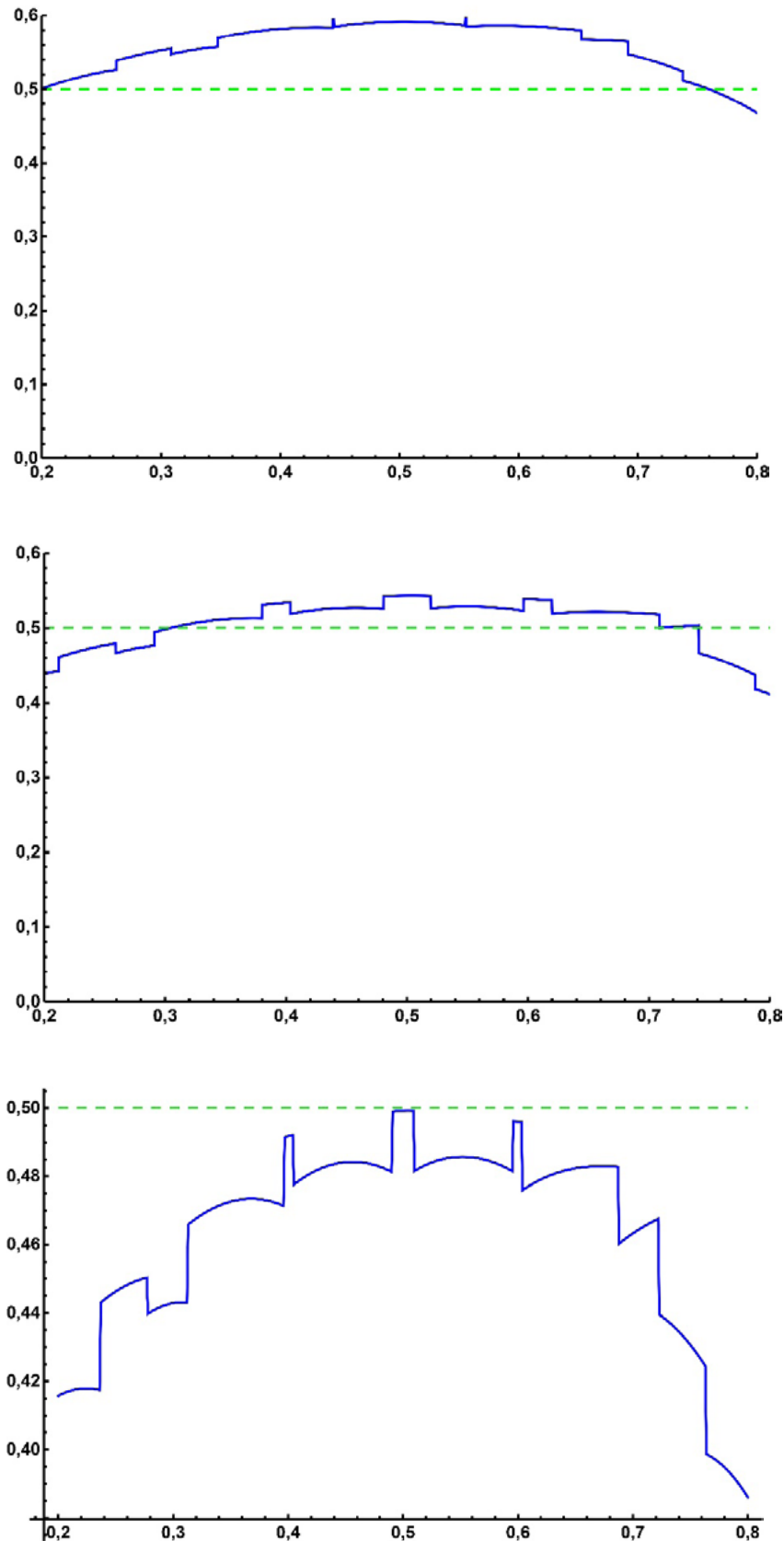


Рис. 11. Графики средней длины доверительных интервалов неизвестного значения параметра биномиального распределения для примера 7. Ось абсцисс – возможные значения вероятности оцениваемого параметра, ось ординат – соответствующая средняя длина. Сверху вниз для доверительных интервалов на основе: обычной функции  $p$ -значения (28), функции среднего  $p$ -значения (29), функции асимптотического  $p$ -значения (30)

Как видно из рис. 11, консервативная функция  $p$ -значения (28) порождает более широкие интервалы, чем две другие функции, причем наименьшие интервалы порождает асимптотическая функция (30), которая, однако, имеет худшую вероятность покрытия (см. рис. 10). Поэтому задача определения «наилучшего» способа определения доверительного интервала на основе понятий вероятности покрытия и средней длины доверительного интервала требует внимательного анализа имеющихся данных, так как эти критерии взаимно дополнительные. Увеличение одной из этих характеристик приводит к увеличению и другой, в то время как требуется увеличение вероятности покрытия при уменьшении длины интервала. Разумной видится следующая стратегия: среди всех рассматриваемых способов вычисления доверительного интервала следует выбрать те, для которых вероятность покрытия будет наибольшей для того диапазона значений, в котором рассматриваемый параметр должен находиться, а среди этих методов выбрать тот, для которого средняя длина интервала будет наименьшей.

**Пример 7 (продолжение).** При описанных в примере 7 условиях ( $n = 10, k = 2$ ) можно предположить, что оцениваемая биномиальная вероятность обнаружения побочных явлений находится в пределах  $(0,1; 0,3)$ . Найдем минимальную вероятность покрытия при  $\pi \in (0,1; 0,3)$  и максимальную среднюю длину доверительных интервалов при этом условии для функций  $p$ -значений (28)–(30) (рис. 12–14).

Таким образом, гарантированную вероятность покрытия дает только консервативный метод (28), который, однако, имеет наиболее широкие доверительные интервалы. Тем не менее, если известно, что значение оцениваемой вероятности не может превышать 0,25, то такую же гарантированную вероятность обеспечивает и функция (29), имеющая при этом заметно меньшую среднюю длину доверительных интервалов. Наиболее точные доверительные интервалы (наименьшей средней длины) обеспечивает асимптотическая функция (30), которая, однако, имеет худшую вероятность покрытия. Для ее корректного применения с гарантированной вероятностью покрытия необходимо ограничиваться теми интервалами возможных значений параметра, на которых вероятность покрытия больше заданного номинального уровня (например,  $(0,18; 0,24)$ ).

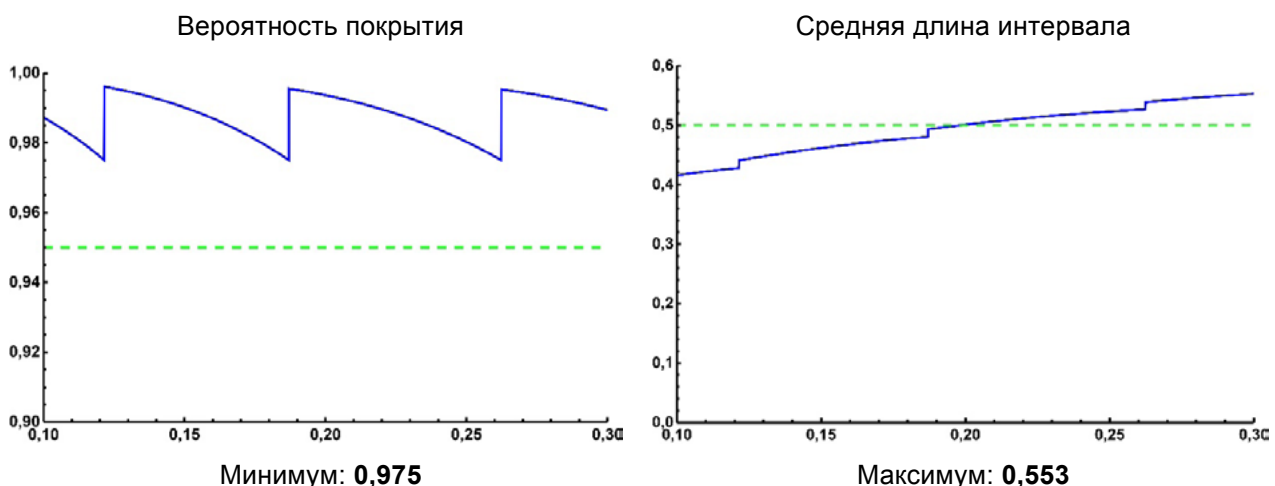


Рис. 12. Вероятность покрытия и средняя длина доверительных интервалов на интервале  $(0,1; 0,3)$  для примера 7 для двусторонней функции  $p$ -значения (28). По оси абсцисс отложены возможные значения вероятности для примера 7, по оси ординат – вероятность покрытия (слева) и средняя длина доверительного интервала (справа)

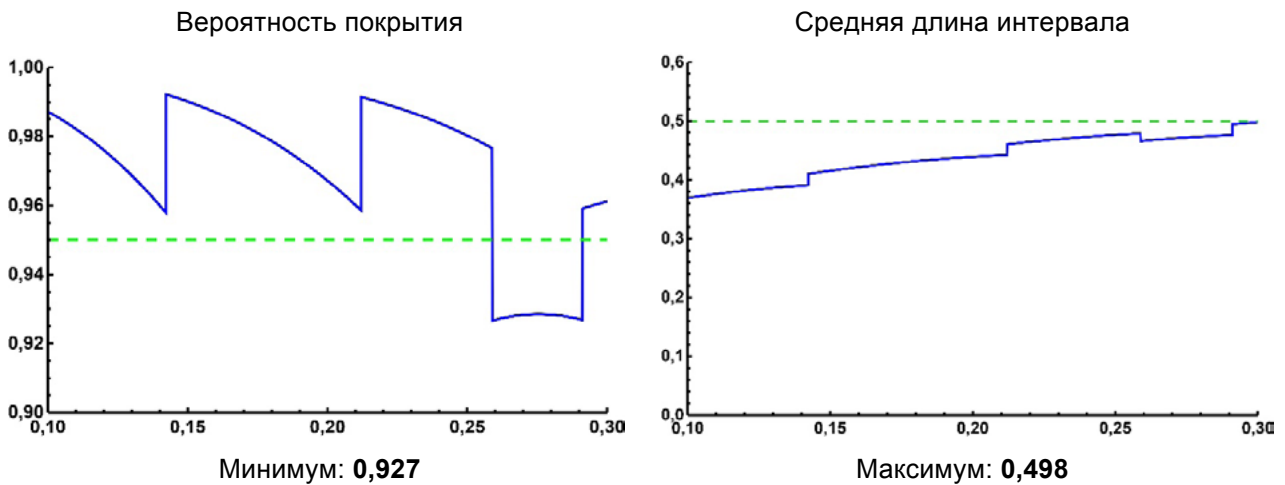


Рис. 13. Вероятность покрытия и средняя длина доверительных интервалов на интервале  $(0,1; 0,3)$  для примера 7 для функции  $p$ -значения (29). По оси абсцисс отложены возможные значения вероятности для примера 7, по оси ординат – вероятность покрытия (слева) и средняя длина доверительного интервала (справа)

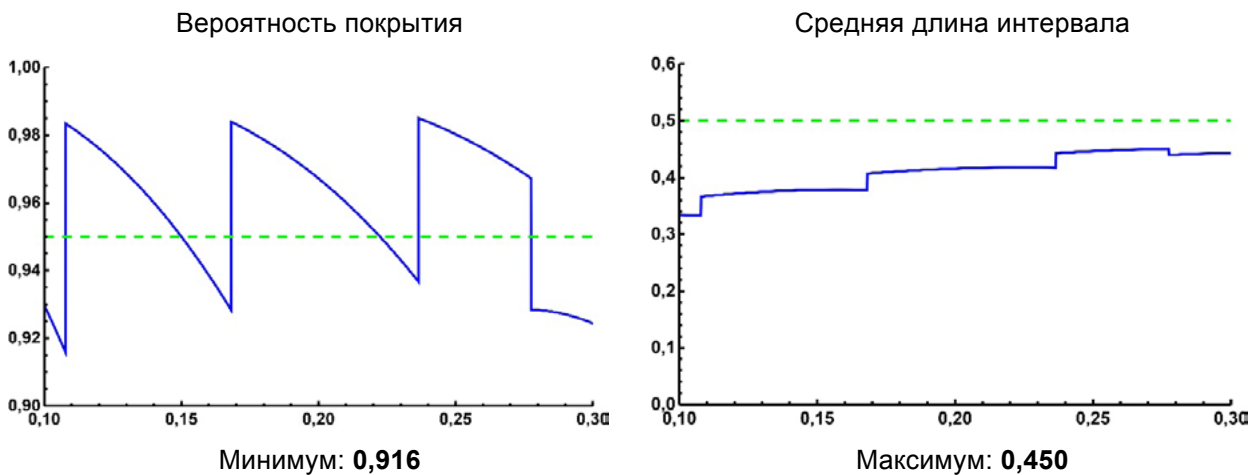


Рис. 14. Вероятность покрытия и средняя длина доверительных интервалов на интервале  $(0,1; 0,3)$  для примера 7 для функции  $p$ -значения (30). По оси абсцисс отложены возможные значения вероятности для примера 7, по оси ординат – вероятность покрытия (слева) и средняя длина доверительного интервала (справа)

## 5. Заключение

1. Анализ дискретных данных представляет специфическую область теоретической и прикладной статистики, имеющую широкие приложения в различных науках. Использование для анализа таких данных методов непрерывных случайных величин дает в лучшем случае приближение, точность которого остается дискуссионной, а чаще всего необсуждаемой. Однако для малых выборок эта точность может оказаться существенно недостаточной для получения статистически значимых результатов. В то же время теоретические исследования способов точного анализа дискретных данных ведутся уже довольно давно, и полученные важные результаты этой работы реализованы в нескольких статистических пакетах общего назначения и некоторых специализированных. Однако грамотное приложение этих методов и применение компьютерных программ требует знания основных понятий точного анализа

данных и понимания особенностей применения соответствующих методов. Несмотря на простоту аналитического аппарата изложенных методов, идеи точного анализа используют фундаментальные понятия теории статистического оценивания и вывода. Их знание необходимо не только при анализе дискретных данных, но и для понимания статистики вообще. Приведенные в работе определения и примеры помогут как в первоначальном знакомстве с этими понятиями, так и в последующем освоении методов точного анализа данных.

## 6. Список литературы

1. *Флетчер, Р.* Клиническая эпидемиология : Основы доказат. медицины / Р. Флетчер, С. Флетчер, Э. Вагнер; пер. с англ. под общ. ред. С. Е. Бащинского и С. Ю. Варшавского; [Пер. с англ. А. Д. Деев и др.]. – 3-е изд. – М. : Медиа Сфера, 2004. – 347 с. – ISBN 5-89084-011-8. – EDN QLFECB.
2. *Гринхальх, Т.* Основы доказательной медицины / Т. Гринхальх ; Триша Гринхальх ; пер. с англ.; под ред. И. Н. Денисова, К. И. Сайткулова. – 3-е изд. – М. : ГЭОТАР-Медиа, 2009. – 282 с. – ISBN 978-5-9704-1347-0. – EDN QLULXN.
3. *Эттингер, А. П.* Что такое доказательная медицина? / А. П. Эттингер, М. Е. Жарова // Доказательная гастроэнтерология. – 2021. – Т. 10, № 1. – С. 38–48. – DOI 10.17116/dokgastro20211001138. – EDN WDYYQS.
4. *Общая эпидемиология с основами доказательной медицины / под ред. В. И. Покровского, Н. И. Брико.* – 2-е изд., испр. и доп. – М. : ГЭОТАР-Медиа, 2012. – 496 с.
5. *Гланц, С.* Медико-биологическая статистика / С. Гланц ; под ред. Н. Е. Бузикашвили и Д. В. Самойлова. – М. : Практика, 1999. – 459 с. – ISBN 5-89816-009-4.
6. *Брико, Н. И.* Эпидемиология : учеб. в 2 т. / Н. И. Брико, Л. П. Зуева, В. И. Покровский [и др.]. – М. : Издательство «Медицинское информационное агентство», 2013. – 832 с. – ISBN 978-5-9986-0110-1. – EDN RTTTRD.
7. *Власов, В. В.* Эпидемиология : учеб. пособие для студентов вузов, обучающихся по специальности 040300 – Медико-профилактическое дело / В. В. Власов. – 2-е изд., испр. – М. : ГЭОТАР-Медиа, 2006. – 462 с. – ISBN 5-9704-0265-6. – EDN QLNFJP.
8. *Rothman, K. J.* Modern epidemiology / K. J. Rothman, S. Greenland, T. L. Lash. – 3<sup>rd</sup> ed., 2011. – 758 p. – EDN YDXWPS.
9. *Schoenbach, V. J.* Understanding the Fundamentals of Epidemiology / V. J. Schoenbach, W. D. Rosamond. – Department of Epidemiology School of Public Health University of North Carolina at Chapel Hill, 2000. – 584 с.
10. *Ahrens, W.* Handbook of Epidemiology / W. Ahrens, I. Pigeot. – Springer, 2005. – 1594 p. – ISBN 3-540-00566-8.
11. *Платонов, А. Е.* Статистический анализ в медицине и биологии: задачи, терминология, логика, компьютерные методы / А. Е. Платонов. – М., 2000. – 52 с. – ISBN 5-7901-0022-8. – EDN PBDIJN.
12. *Малета, Ю. С.* Математические методы статистического анализа в биологии и медицине / Ю. С. Малета, В. В. Тарасов. – М. : Изд-во МГУ, 1982. – 179 с.
13. *Quinn, G. P.* Experimental Design and Data Analysis for Biologists / G. P. Quinn, M. J. Keough. – Cambridge University Press, 2002. – 557 p. – ISBN 0-521-00976-6.
14. *Любищев, А. А.* Дисперсионный анализ в биологии / А. А. Любищев – М. : МГУ, 1986. – 93 с.



15. Mathematics for Ecology and Environmental Sciences / под ред. Y. Takeuchi, Y. Iwasa, K. Sato. – Springer, 2007. – 186 p. – ISBN 10 3-540-34427-6.
16. Legendre, P. Numerical Ecology / P. Legendre, L. Legendre. – Elsevier, 1998. – 870 p. – ISBN 0-444-89250-8.
17. Магнус, Я. Р. Эконометрика. Начальный курс : учеб. / Я. Р. Магнус, П. К. Катыхов, А. А. Пересецкий. – М. : Издательский дом «Дело» РАНХиГС, 2021. – 504 с. – ISBN 978-5-85006-296-5. – EDN MDSVQL.
18. Мхитарян, В. С. Эконометрика : учеб. пособие / В. С. Мхитарян, М. Ю. Архипова, В. П. Сиротин. – М. : Евразийский открытый институт, 2012. – 224 с. – ISBN 978-5-374-00053-5. – EDN RBAOLL.
19. Джонстон, Дж. Эконометрические методы / Дж. Джонстон. – М. : Статистика, 1980. – 440 с.
20. Handbook of Computational Econometrics / под ред. D. A. Belsley, E. J. Kontoghiorghes. – Wiley, 2009. – 516 p. – ISBN 978-0-470-74385-0.
21. Грин, У. Г. Эконометрический анализ / У. Г. Грин; пер. с англ. под науч. ред. С. С. Синельникова и М. Ю. Турунцевой. – М. : Изд. дом «Дело», 2016. Кн. 1. – 760 с. Кн. 2 – 1476 с. – ISBN 978-5-7749-1158-5.
22. Феллер, В. Введение в теорию вероятностей и ее приложения = An Introduction to probability theory and its applications / В. Феллер; пер. со второго англ. изд. и предисл. Ю. В. Прохорова. – 2-е изд. – М. : URSS, 2009. – 751 с. – ISBN 978-5-397-01036-8. – EDN QJVXJV.
23. Боровков, А. А. Теория вероятностей : учеб. пособие для студентов высших учебных заведений, обучающихся по направлению подготовки 010100 «Математика» / А. А. Боровков. – 5-е изд., существенно перераб. и доп. – М. : URSS, 2009. – 652 с. – ISBN 978-5-397-00582-1. – EDN QJVLHP.
24. Гаек, Я. Теория ранговых критериев / Я. Гаек, З. Шидак. – М. : Наука, 1971. – 376 с.
25. Холлендер, М. Непараметрические методы статистики / М. Холлендер, Д. Вульф. – М. : Финансы и статистика, 1983. – 518 с.
26. Corder, G. W. Nonparametric statistics for non-statisticians. A Step-by-Step Approach / G. W. Corder, D. I. Foreman. – Wiley, 2009. – 264 p. – ISBN 978-0-470-45461-9.
27. Fisher, R. A. Statistical Methods for Research Workers. – 5<sup>th</sup> ed. / R. A. Fisher. – Oliver and Boyd, Edinburgh, 1934. – 336 p.
28. Fisher, R. A. The logic of inductive inference (with discussion) / R. A. Fisher // J. Royal Stat. Soc. – 1935. – Vol. 98. – Part I: 35–82.
29. Irwin, J. O. Tests of significance for differences between percentages based on small numbers / J. O. Irwin // Metron. – 1935. – Vol. 12: 83–94.
30. Yates, F. Contingency tables involving small numbers and the chi-square test / F. Yates // Supplement to the J. Royal Stat Soc, Series B. – 1934. – 1: 217–235.
31. Cytel Software Corporation. StatXact for Windows, version 6, Cytel Software Corporation, Cambridge, MA.
32. Agresti, A. Categorical data analysis / A. Agresti. – Wiley, 2002. – 732 p. – ISBN 0-471-36093-7.
33. Sakamoto, Y. Categorical data analysis by AIC / Y. Sakamoto // Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach: Vol. 2. Multivariate Statistical Modeling. 1994. – 255–269. – ISBN 978-94-011-0800-3. – DOI: 10.1007/978-94-011-0800-3\_10.

34. *Birnbaum, A.* Confidence curves: an omnibus technique for estimation and testing statistical hypotheses / A. Birnbaum // *J. Am Stat Assoc.* – 1961. – Vol. 56. – С. 246–249. – DOI:10.1080/01621459.1961.10482107.
35. *Cox, D. R.* Analysis of binary data / D. R. Cox, E. J. Snell. – London, New York. Chapman and Hall, 1989. – 247 p. – ISBN 0-412-30620-4.
36. *Zelen, M.* The analysis of several 2×2 contingency tables / M. Zelen // *Biometrika.* – 1971. – Vol. 58. – No. 1. – P. 129. – EDN ILOSAN.
37. *Emerson, J. D.* Combining estimates of the odds ratio: the state of the art / J. D. Emerson // *Stat Methods in Med Research.* – 1994. – Vol. 3. – P. 157–178. – DOI: 10.1177/09622809400300204.
38. *Hirji, K. F.* Exact Analysis of Discrete Data / K. F. Hirji. – Chapman & Hall/CRC. – 2006. – 521 p. – ISBN 1-58488-070-8.
39. *May, R. M.* Host-Parasitoid Systems in Patchy Environments: A Phenomenological Model / R. M. May // *J. Animal Ecology.* – 1978. – Vol. 47, N 3. – P. 833–844. – DOI: 10.2307/3674.
40. *Ross, G. J. S.* The Negative Binomial Distribution / G. J. S. Ross, D. A. Preece // *J. Royal Stat Soc. Series D.* – 1985. – Vol. 34. – N 3. – С. 323–335. – DOI: 10.2307/2987659.
41. *Amrhein, L.* A mechanistic model for the negative binomial distribution of single-cell mRNA counts / L. Amrhein, K. Harsha, K. Fuchs // *bioRxiv.* – 2019. – DOI: 10.1101/657619.
42. *Johnson, N. L.* Univariate Discrete Distributions / N. L. Johnson, A. W. Kemp, S. Kotz // John Wiley & Sons, Inc., Hoboken, NJ, 2005. – 677 p. – ISBN 0-471-27246-9.
43. *Tuwei, K. E.* Power series distributions / K. E. Tuwei // [http://erepository.uonbi.ac.ke/bitstream/handle/11295/71878/Tuwei\\_Power series distribution.pdf](http://erepository.uonbi.ac.ke/bitstream/handle/11295/71878/Tuwei_Power%20series%20distribution.pdf).
44. *Rubin-Delanchy, P.* Meta-analysis of mid-p-values: some new results based on the convex order / P. Rubin-Delanchy, N. A. Heard, D. J. Lawson // *J. Am Stat Assoc.* – 2018. – Vol. 114(527). – P. 1105–1112. – DOI: 10.1080/01621459.2018.1469994.
45. *Moyé, L.* Statistical reasoning in Medicine: The intuitive *p*-value primer / L. Moyé – Springer-Verlag, NY, 2006. – 301 p. – ISBN 0-387-32913-7.
46. *Леман, Э.* Проверка статистических гипотез / Э. Леман. – М. : Наука, 1979. – 409 с.
47. *Casella, G.* Statistical inference / G. Casella, R. L. Berger. – Duxbury, 2002. – 660 p. – ISBN 0-534-23412-6. – EDN QJRALL.

## Сведения об авторах:

**Панов Владимир Григорьевич**, к. ф.-м. н., доцент, ведущий научный сотрудник лаборатории математического моделирования в экологии и медицине Института промышленной экологии УрО РАН, ул. С. Ковалевской, 20, г. Екатеринбург, Россия. Эл. почта: [vpanov@ecko.uran.ru](mailto:vpanov@ecko.uran.ru)

# ANALYSIS OF DISCRETE DATA. BASIC NOTIONS

V. G. Panov

*Institute of Industrial Ecology of UrB of RAS, Ekaterinburg, Russia*

*The review article is devoted to methods of discrete data analysis. The main attention is paid to exact methods based on the representation of discrete probability distributions as power series distributions. The examples presented in the article relate to medicine and biology.*

**Key words:** discrete probability distributions; power series distributions; exact data analysis methods; statistical hypotheses; confidence regions; power of a criterion; coverage probability.

## References

1. *Fletcher, R.* Clinical epidemiology: The basics of evidence-based medicine / R. H. Fletcher; S. W. Fletcher, E. Wagner; Translation from English, Eds. S. E. Bashchinsky and C. Y. Varshavsky; [Trans. Eng. A. D. Deev and others]. – 3<sup>rd</sup> ed. – Moskow : Media Spere, 2004. – 347 p. – ISBN 5-89084-011-8. – EDN QLFECB.
2. *Greenhalgh, T.* The Basics of Evidence-Based Medicine and Healthcare: Translation from English / T. Greenhalgh; Trisha Greenhalgh; Translation Eds. I. N. Denisov, K. I. Saitkulov. – 3<sup>rd</sup> ed. – Moskow : GEOTAR-Media, 2009. – 282 p. – ISBN 978-5-9704-1347-0. – EDN QLULXN.
3. *Ettinger, A. P.* What is the evidence-based medicine? / A. P. Ettinger, M. E. Zharova // Russian J. of Evidence-Based Gastroenterology. – 2021. – Vol. 10, № 1. – P. 38–48. – DOI 10.17116/dokgastro20211001138. – EDN WDYYQS.
4. General Epidemiology with Fundamentals of Evidence-Based Medicine / Eds: V. I. Pokrovsky, N. I. Briko. – 2<sup>nd</sup> ed. – M.: GEOTAR-Media, 2012. – 496 p.
5. *Glantz, S.* Primer of biostatistics / S. Glantz. Trans. Under Eds N. E. Buzikashvili and D. V. Samoilov. – M. : Praktika, 1999. – 459 p. – ISBN 5-89816-009-4.
6. *Briko, N. I.* Epidemiology: Textbook / N. I. Briko, L. P. Zueva, V. I. Pokrovsky [et al.]. – Moskow : Publisher «Medical information agency», 2013. – 832 p. – ISBN 978-5-9986-0110-1. – EDN RTTTRD.
7. *Vlasov, V. V.* Epidemiology / V. V. Vlasov. – 2<sup>nd</sup> ed. – Moskow : GEOTAR-Media, 2006. – 462 p. – ISBN 5-9704-0265-6. – EDN QLNFJP.
8. *Rothman, K. J.* Modern epidemiology / K. J. Rothman, S. Greenland, T. L. Lash. – 3<sup>rd</sup> ed., 2011. – P. 1–758. – EDN YDXWPS.
9. *Schoenbach, V. J.* Understanding the Fundamentals of Epidemiology / V. J. Schoenbach, W. D. Rosamond. – Department of Epidemiology School of Public Health University of North Carolina at Chapel Hill, 2000. – 584 c.

10. *Ahrens, W.* Handbook of Epidemiology / W. Ahrens, I. Pigeot. – Springer, 2005. – 1594 p. – ISBN 3-540-00566-8.
11. *Platonov, A. E.* Statistical Analysis in Medicine and Biology: Problems, Terminology, Logic, Computer Methods / A. E. Platonov. – Moskow, 2000. – 52 p. – ISBN 5-7901-0022-8. – EDN PBDIJN.
12. *Maleta, Y. S.* Mathematical Methods of Statistical Analysis in Biology and Medicine / Y. S. Maleta, V. V. Tarasov. – M. : Moskow State University, 1982. – 179 p.
13. *Quinn, G. P.* Experimental Design and Data Analysis for Biologists / G. P. Quinn, M. J. Keough. – Cambridge University Press, 2002. – 557 p. – ISBN 0-521-00976-6.
14. *Lyubishchev, A. A.* Analysis of variance in biology / A. A. Lyubishchev. – M. : MGU, 1986. – 93 p.
15. *Mathematics for Ecology and Environmental Sciences* / Eds: Y. Takeuchi, Y. Iwasa, K. Sato. – Springer, 2007. – 186 p. – ISBN 10 3-540-34427-6.
16. *Legendre, P.* Numerical Ecology / P. Legendre, L. Legendre. – Elsevier, 1998. – 870 p. – ISBN 0-444-89250-8.
17. *Magnus, Y. R.* Econometrics. Primary course: textbook / Y. R. Magnus, P. K. Katyshev, A. A. Peresetsky. – Moskow: Publishing House «Delo» RANEPА, 2021. – 504 p. – ISBN 978-5-85006-296-5. – EDN MDSVQL.
18. *Mkhitaryan, V. S.* Econometrics : Tutorial / V. S. Mkhitaryan, M. Yu. – Moscow : Eurasian Open Institute, 2012. – 224 p. – ISBN 978-5-374-00053-5. – EDN RBAOLL.
19. *Johnston, J.* Econometric methods / J. Johnston. – M. : Statistika, 1980. – 440 p.
20. *Handbook of Computational Econometrics* / Eds. D. A. Belsley, E. J. Kontoghiorghes. – Wiley, 2009. – 516 p. – ISBN 978-0-470-74385-0.
21. *Greene, W. H.* Econometric analysis / W. H. Greene; Trans. from Eng. under Eds S. S. Sinelnikov and M. Y. Turuntseva. – M. : Publishing House «Delo», 2016. Vol. 1 – 760 p. Vol. 2 – 1476 p. – ISBN 978-5-7749-1158-5.
22. *Feller, W.* An Introduction to probability theory and its applications / W. Feller; Trans. of 2<sup>nd</sup> Eng. ed. – Moskow : URSS, 2009. – 751 p. – ISBN 978-5-397-01036-8. – EDN QJVXJV.
23. *Borovkov, A. A.* Probability theory / A. A. Borovkov. – 5<sup>th</sup> ed. – Moskow : URSS, 2009. – 652 p. – ISBN 978-5-397-00582-1. – EDN QJVLHP.
24. *Hajek, J.* Theory of rank tests / J. Hajek, Z. Shidak. – M. : Nauka, 1971. – 376 p.
25. *Hollander, M.* Nonparametric statistical methods / M. Hollander, D. Wolfe. – M. : Finance and Statistics, 1983. – 518 p.
26. *Corder, G. W.* Nonparametric statistics for non-statisticians. A Step-by-Step Approach / G. W. Corder, D. I. Foreman. – Wiley, 2009. – 264 p. – ISBN 978-0-470-45461-9.
27. *Fisher, R. A.* Statistical Methods for Research Workers. – 5<sup>th</sup> ed. / R. A. Fisher. – Oliver and Boyd, Edinburgh, 1934. – 336 c.
28. *Fisher, R. A.* The logic of inductive inference (with discussion) / R. A. Fisher // J. Royal Stat. Soc. – 1935. – Vol. 98. – Part I: 35–82.
29. *Irwin, J. O.* Tests of significance for differences between percentages based on small numbers / J. O. Irwin // Metron. – 1935. – Vol. 12: 83–94.
30. *Yates, F.* Contingency tables involving small numbers and the chi-square test / F. Yates // Supplement to the J. Royal Stat Soc, Series B. – 1934. – 1: 217–235.
31. Cytel Software Corporation. StatXact for Windows, version 6, Cytel Software Corporation, Cambridge, MA.

32. *Agresti, A.* Categorical data analysis / A. Agresti. – Wiley, 2002. – 732 p. – ISBN 0-471-36093-7.
33. *Sakamoto, Y.* Categorical data analysis by AIC / Y. Sakamoto // Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach: Vol. 2. Multivariate Statistical Modeling. 1994. – 255–269. – ISBN 978-94-011-0800-3. – DOI: 10.1007/978-94-011-0800-3\_10.
34. *Birnbaum, A.* Confidence curves: an omnibus technique for estimation and testing statistical hypotheses / A. Birnbaum // J. Am Stat Assoc. – 1961. – Vol. 56. – C. 246–249. – DOI:10.1080/01621459.1961.10482107.
35. *Cox, D. R.* Analysis of binary data / D. R. Cox, E. J. Snell. – London, New York. Chapman and Hall. – 1989. – 247 c. – ISBN 0-412-30620-4.
36. *Zelen, M.* The analysis of several 2×2 contingency tables / M. Zelen // Biometrika. – 1971. – Vol. 58, No. 1. – P. 129. – EDN ILOSAN.
37. *Emerson, J. D.* Combining estimates of the odds ratio: the state of the art / J. D. Emerson // Stat Methods in Med Research. – 1994. – Vol. 3. – C. 157–178. – DOI: 10.1177/096228029400300204.
38. *Hirji, K. F.* Exact Analysis of Discrete Data / K. F. Hirji. – Chapman & Hall/CRC. – 2006. – 521 c. – ISBN 1-58488-070-8.
39. *May, R. M.* Host-Parasitoid Systems in Patchy Environments: A Phenomenological Model / R. M. May // J Animal Ecology. – 1978. – Vol. 47, N 3. – C. 833–844. – DOI: 10.2307/3674.
40. *Ross, G. J. S.* The Negative Binomial Distribution / G. J. S. Ross, D. A. Preece // J. Royal Stat Soc. Series D. – 1985. – Vol. 34, N 3. – C. 323–335. – DOI: 10.2307/2987659.
41. *Amrhein, L.* A mechanistic model for the negative binomial distribution of single-cell mRNA counts / L. Amrhein, K. Harsha, K. Fuchs // bioRxiv. – 2019. – DOI: 10.1101/657619.
42. *Johnson, N. L.* Univariate Discrete Distributions / N. L. Johnson, A. W. Kemp, S. Kotz // John Wiley & Sons, Inc., Hoboken, NJ, 2005. – 677 c. – ISBN 0-471-27246-9.
43. *Tuwei, K. E.* Power series distributions / K. E. Tuwei // [http://erepository.uonbi.ac.ke/bitstream/handle/11295/71878/Tuwei\\_Power series distribution.pdf](http://erepository.uonbi.ac.ke/bitstream/handle/11295/71878/Tuwei_Power%20series%20distribution.pdf).
44. *Rubin-Delanchy, P.* Meta-analysis of mid-p-values: some new results based on the convex order / P. Rubin-Delanchy, N. A. Heard, D. J. Lawson // J. Am Stat Assoc. – 2018. – Vol. 114(527). – P. 1105–1112. – DOI: 10.1080/01621459.2018.1469994.
45. *Moyé, L.* Statistical reasoning in Medicine: The intuitive *p*-value primer / L. Moyé – Springer-Verlag, NY, 2006. – 301 c. – ISBN 0-387-32913-7.
46. *Lehmann, E.* Testing statistical hypotheses / E. Lehmann. – M. : Nauka, 1979. – 409 p.
47. *Casella, G.* Statistical inference / G. Casella, R. L. Berger. – Duxbury, 2002. – 660 c. – ISBN 0-534-23412-6. – EDN QJRALL

## **ПРЕДИСЛОВИЕ К СТАТЬЕ В. Н. ЧУКАНОВА «ЭКОЛОГИЧЕСКИЕ ПРОБЛЕМЫ – СИСТЕМНЫЙ ПОДХОД»**

Статья основателя Института промышленной экологии УрО РАН Виктора Николаевича Чуканова, написанная им в 1999 г., подводящая промежуточные итоги деятельности Института, оказалась не только научной, но также программной статьей, формулирующей задачи на будущее. Необходимость комплексного, системного подхода к решению задач, касающихся, на первый взгляд, только проблем загрязнения окружающей среды, была актуальна в 1999 г. и продолжает быть актуальной по сей день.

В начале 90-х гг., когда формировался Институт промышленной экологии, было много желающих списать все беды новой России на загрязнение окружающей среды (как часто говорят, на «плохую экологию» – совершенно бессмысленный термин). В качестве примера такого подхода можно назвать монографии тех лет, написанные Ю. П. Гичевым. Была ли это искренняя позиция научных работников или желание угодить новой власти – сейчас уже не узнать, но это было четкое послание российскому обществу: все беды идут от промышленности, а руководители страны ни в чем не виноваты. В то же время были попытки объяснить упадочное состояние Российского общества социально-экономическими проблемами страны. Достаточно вспомнить работы академика Б. Т. Величковского, который на обширном научном материале убедительно показал сильнейшее влияние экономического состояния страны на все стороны жизни в России (в том числе на заболеваемость и смертность населения).

Предлагаемый Виктором Николаевичем комплексный подход к научному изучению экологических проблем – это призыв рассмотреть проблему со всех сторон, подход, который должен ответить на экологические вопросы наиболее полно и адекватно. Надеюсь, что статья В. Н. Чуканова «Экологические проблемы – системный подход» будет интересна и полезна современным исследователям проблем нашего общества.

А. Н. Варакин, д. физ.-мат. наук, профессор,  
г. н. с. ИПЭ УрО РАН

# ЭКОЛОГИЧЕСКИЕ ПРОБЛЕМЫ – СИСТЕМНЫЙ ПОДХОД

В. Н. Чуканов

*Институт промышленной экологии УрО РАН, г. Екатеринбург, Россия*

## **Введение**

Институт промышленной экологии УрО РАН организован в 1992 г. на базе Научно-инженерного центра экологической безопасности УрО РАН.

Создание и становление Института совпало с тяжелым переделом в жизни нашей страны. Это не могло не отразиться на выборе научной тематики и характере решаемых задач. Необходимо было выйти на качественно иной уровень обобщения, чтобы увидеть место экологических проблем в ряду всех остальных: медицинских, социальных, экономических, демографических и т. д., причем увидеть во взаимосвязи и взаимозависимости. Первый шаг в понимании этого сделан.

Основной вывод – эффективное решение масштабных многофакторных и разноплановых экологообусловленных проблем возможно только при комплексном системном подходе.

Эффективно решить проблемы по отдельности: экологические, основываясь только на экологических принципах; экономические, учитывая только экономические аспекты; медицинские, рассматривая только проблемы здоровья населения и т. д., невозможно. Связано это с тем, что, как правило, экологические проблемы решаются в комплексе в целях, например, улучшения здоровья населения, экономической или демографической сатурации, повышения качества жизни и т. д. Но когда мы увязываем состояние окружающей природной среды и здоровья населения, окружающей среды и экономики, начинают действовать свои критерии и принципы, которые отличаются от чисто экологических, медицинских, экономических принципов, например, определения приоритетов.

Задача, стоящая сегодня перед коллективом Института, – на основе принципов системного анализа разработать методологию решения комплексных и экологозависимых проблем на уровнях от локального, регионального и выше.

Это и определило основные научные направления:

- Системный анализ и моделирование многофакторного воздействия на биологические среды и здоровье населения; экологическая безопасность; риски.
- Устойчивое развитие территорий – системный анализ экологических, социальных и экономических проблем и оценка соответствующих последствий при различных сценариях развития территории.
- Исследование процессов тепломассопереноса, характерных для проблем экологии.
- Приведем ряд результатов применения системного подхода, полученных сотрудниками Института.

## **Проблема радиоактивно-загрязненных территорий Уральского региона – системный подход**

По совокупности проблем, сложившихся в Уральском регионе, радиационно-экологическая ситуация не имеет аналогов. Проблемы связаны в основном с последствиями деятельности первого в нашей стране атомно-промышленного комплекса по наработке оружейного плутония (ПО «Маяк» Челябинской области). Их условно можно разделить на два класса: проблемы накопленных отходов и проблемы, связанные с загрязнением окружающей среды.

На промплощадке ПО «Маяк» накоплено более 1 млрд Ки активности жидких и твердых радиоактивных отходов. Значительная часть их депонирована в открытых природных средах. Особую тревогу вызывает оз. Карачай, в котором находится около 120 млн Ки, и каскад радиоактивных водоемов на р. Тече. Под озером образовалась линза радиоактивно-загрязненных подземных вод, которая движется вдоль меридионального разлома. В южном направлении она уже прошла под ложем р. Мишеляка, притока р. Течи, в которую отмечены выходы радиоактивности.

Загрязнение региона обусловлено газоаэрозольными выбросами, особенно в начальный период работы, в результате штатной деятельности предприятия и тремя радиационными инцидентами:

Сброс радиоактивных отходов суммарной активностью около 2,75 млн Ки в р. Течу, в основном в период с 1949 по 1952 г., что привело к загрязнению гидрографической системы Теча – Исеть – Тобол – Иртыш – Обь.

Тепловой взрыв емкости с высокоактивными жидкими отходами 29.09.57. Около 2 млн Ки активности было поднято в воздух, и радиоактивное облако прошло в сторону Тюмени, образовав Восточно-Уральский радиоактивный след (ВУРС). В зону ВУРСа попали десятки населенных пунктов, в том числе города Багаряк, Каменск-Уральский, Камышлов.

Разнос радиоактивной пыли активностью ~600 Ки (апрель 1967 г.) с обнажившейся донной поверхности оз. Карачай.

Разработка комплексов научных рекомендаций по оздоровлению радиоактивно-экологической ситуации была начата в 1990 г. в рамках специально созданных комиссий АН СССР и президента СССР. Основные итоговые документы этих комиссий сформулированы группой специалистов ИПЭ (в то время НИЦЭБ) УрО РАН. В них содержались предложения о Государственной программе по реализации срочных мер, аналогичной Чернобыльской, обоснованы ее направления и комплексная структура. На всех последующих этапах разработки, научного обеспечения и реализации программных мер Институт принимал непосредственное участие. При подготовке Государственной программы РФ по радиационной реабилитации Уральского региона и мерах оказания помощи пострадавшему населению на период 1992–1995 гг. ИПЭ от имени УрО РАН разрабатывал ее научный раздел. В период реализации Программы Институт обеспечивал координацию научных исследований.

К сожалению, экономические условия в стране не позволили реализовать Программу в полном объеме. В результате при участии Института была разработана и утверждена новая Федеральная целевая программа на период до 2000 г. В настоящее время, опять же при участии Института, разрабатывается программа социальной реабилитации и радиационной безопасности региона на период до 2010 г.

За минувшие восемь лет под научным руководством Института и при его непосредственном участии выполнены комплексные исследования долгосрочных



последствий радиоактивного загрязнения для окружающей среды, здоровья населения, социально-экономической сферы пострадавших территорий. Получены результаты фундаментального характера, заставившие принципиально пересмотреть концепцию реабилитации населения и территории.

На основе системного анализа установлено, что:

- загрязнение носит фрактальный характер; начальная версия радиоактивного загрязнения территории, заложенная в первой Программе, не соответствует действительности;
- полученная оценка накопленной коллективной дозы населением  $\sim 5\text{--}104\text{ Зв}$  позволяет сделать вывод, что суммарное радиационное воздействие, обусловленное деятельностью ПО «Маяк», сопоставимо с последствиями Чернобыльской катастрофы;
- радиационный фактор, приоритетный с экологической точки зрения, не является приоритетным или определяющим в ряду факторов, формирующих здоровье жителей г. Каменска-Уральского. Однако иммунно-гематологический статус населения, проживающего с момента аварии на радиоактивно-загрязненной территории, в том числе потомков во втором и третьем поколениях, отличается от такового для населения, проживающего на радиоактивно чистой территории;
- выявленный масштаб опосредованных последствий радиационных аварий – снижение уровня и качества жизни, ухудшение общего состояния здоровья и депопуляция населения – привел к ущербу, на порядок превышающему необходимые затраты на реабилитацию населения и территорий;
- в настоящее время в регионе развивается крупномасштабная авария, обусловленная неуправляемым распространением радионуклидов в гидросфере.

### **Применимость методов системного анализа**

С самого начала принципиально важно было определить степень применимости методов системного анализа для решения взаимозависимых экологообусловленных проблем. Например, насколько обоснованно можно применять один из мощнейших методов системного анализа – метод анализа иерархий.

Одна аксиома этого метода гласит, что при объединении нескольких систем в одну, т. е. при выходе на более высокий уровень обобщения, или уровень иерархии, приоритеты нового уровня не зависят от приоритетов более низкого уровня. Системные и масштабные исследования по проверке этого принципа проведены в г. Каменске-Уральском и Каменском районе Свердловской области.

В результате аварии на ПО «Маяк» северная часть г. Каменска-Уральского оказалась загрязненной радионуклидом  $^{90}\text{Sr}$  до уровней, в среднем превышающих  $4\text{ Ки/км}^2$ . Критерием безусловного отселения жителей с радиоактивно-загрязненных территорий был уровень загрязнения  $2\text{ Ки/км}^2$ . Период полураспада  $^{90}\text{Sr}$  около 30 лет. Поэтому, по сути дела, население этой части города в минувшие сорок лет жило в условиях, соответствующих статусу зоны экологического бедствия по радиационному фактору. Ни по какому другому фактору – состоянию атмосферы, почвы, гидросферы – положение не соответствовало и не соответствует этому статусу. Оставаясь на экологических принципах, необходимо признать радиационный фактор приоритетным и, соответственно, программа реабилитации территории в первую очередь должна быть направлена именно на его нейтрализацию. Логично предположить, что этот фактор будет определяющим среди факторов, формирующих

здоровье населения, особенно если рассматривать онкологические заболевания, поскольку основным следствием радиационного воздействия являются раковые заболевания.

Однако исследования показали, что радиационный фактор не является определяющим даже для раковых заболеваний. Только для рака молочной железы он занимает пятое ранговое место, для остальных локализаций – еще менее значимые места. Среди факторов, формирующих здоровье детей дошкольного возраста, радиационный занимает ранговые места с седьмого и ниже, соответственно, вклад его в общее изменение здоровья не превышает нескольких процентов.

Следовательно, программа реабилитации, составленная только на основе экологических принципов, будет неэффективна. Таких примеров можно привести достаточно много. Можно сделать вывод, что для такого уровня и масштаба задач применение метода анализа иерархий оправданно. Применение же принципиальной схемы системного анализа возможно к любой проблеме.

### Технология системного анализа

Покажем технологию применения системного анализа на примере выделения или оценки вклада радиационного фактора в формирование здоровья населения, в частности, здоровья детей.

Общая схема анализа такова:

Реализация глобальной цели методами системного анализа



1. Определяется глобальная цель. В нашем случае – это оценка влияния хронического облучения, связанного с длительным проживанием на радиоактивно-загрязненных территориях, на здоровье населения.

2. Глобальная цель разбивается на дерево, или иерархию целей, до уровня, где используются стандартные методы достижения цели.

Первый иерархический уровень целей составляют общие факторы, способные оказывать влияние на здоровье населения, а именно загрязнение окружающей среды и социально-экономические факторы. К этому уровню относится и оценка параметров здоровья населения.

Второй уровень представляет собой детализацию общих факторов. Оценка состояния окружающей среды разделяется на два блока: на оценку радиоактивного и общетехногенного загрязнения. Социально-экономические факторы подразделяются на уровень и качество жизни и демографические факторы.

Третий уровень еще более конкретизирован. В основном на этом уровне производятся измерение и расчет составляющих общих факторов или параметров и ранжирование по ним территории. В нашем случае третий уровень выглядит следующим образом.

Оценка радиоактивного загрязнения территории Восточно-Уральского радиоактивного следа (ВУРС):

1. Оценка радиоактивного загрязнения территории ВУРСа на настоящий момент.
2. Реконструкция первоначального характера и уровней загрязнения зоны ВУРСа.
3. Разработка методов расчетных и экспериментальных ретроспективных оценок коллективных и индивидуальных доз облучения.
4. Классификация населения по уровням облучения.

Оценка общетехногенного загрязнения:

1. Экологическая характеристика источников загрязнения окружающей среды – количественный и качественный состав выбросов и сбросов загрязняющих веществ (металлы, органические и неорганические соединения).
2. Загрязнение атмосферного воздуха – расчет полей концентраций загрязняющих веществ в зависимости от высоты и метеорологических условий (максимальных разовых концентраций, усредненных за различные периоды усреднения).
3. Загрязнение почвы, снега, подземных и поверхностных вод, продуктов питания, питьевой воды.

Оценка уровня жизни населения:

1. Средняя заработная плата.
2. Выплаты социального характера.
3. Прожиточный минимум на душу населения в месяц.
4. Уровень жизни населения  $[(1)+(2)]/(3)$ .
5. Безработица, учтенная и скрытая.
6. Покупательная способность населения.
7. Уровень потребления продуктов питания в среднем за месяц на 1 члена семьи.
8. Розничный товароборот.

Когда речь идет о здоровье детей, в дополнение к характеристике уровня жизни дается характеристика детских дошкольных учреждений, включающая в себя 19 параметров.

Оценка качества жизни населения:

1. Обеспеченность жильем, м<sup>2</sup> на чел.
2. Благоустройство жилого фонда.

3. Обеспеченность больницами и поликлиниками.
4. Обеспеченность школами и детскими дошкольными учреждениями.
5. Состояние социальной инфраструктуры: автомобильные дороги, телефонная сеть, обеспеченность культурными учреждениями и т. д.

Привести все или значительную часть результатов исследований по блокам 3-го уровня в данной статье не представляется возможным. Они опубликованы в ряде статей и монографий. Приведем лишь два примера, характеризующих масштаб и сложность проблем.

Для установления характера текущего загрязнения территории ВУРСа  $^{90}\text{Sr}$  только в Свердловской области обследована территория в 1 600 км<sup>2</sup>; отобрано около 3 000 проб из почти 440 точек отбора. На основе полученных материалов созданы почти рабочие карты-схемы плотности загрязнения территории: Каменского района, М 1:100 000; Богдановичского района, М 1:100 000; г. Каменска-Уральского, М 1:50 000; г. Камышлова, М 1:25 000.

Для восстановления первоначального уровня загрязнения было проведено исследование вертикальной миграции радионуклидов в почве и определен период ее полуочищения. Характер распределения содержания  $^{90}\text{Sr}$  по высоте практически одинаков для почв данного региона (чернозем выщелоченный, почва серая лесная и дерново-подзолистая) и имеет деформированную колоколообразную форму. Спустя 40 лет после аварии 70–90 % от общего на данный момент содержания радионуклида находятся в поверхностном корнеобитаемом слое толщиной 15–20 см.

Определение накопленной дозы и распределение населения по величине накопленной дозы, опять же только для Свердловской области, было выполнено для 86 населенных пунктов. Источниками облучения населения являются загрязненные территории в результате аварии и штатные газоаэрозольные выбросы.

Оценка доз проводилась с учетом внешнего облучения, а также поступления радионуклидов с пищей и ингаляционным путем. Наибольшие трудности возникли при расчете доз в начальный период после аварии, когда наряду с  $^{90}\text{Sr}$  территория была загрязнена относительно короткоживущими излучателями. Ориентировочный состав радиоактивного выброса  $^{90}\text{Sr} + ^{90}\text{Y} - 5,4$ ;  $^{95}\text{Zr} + ^{95}\text{Nb} - 24,9$ ;  $^{144}\text{Ce} + ^{144}\text{Pr} - 66$ ;  $^{106}\text{Ru} + ^{106}\text{Rh} - 3,7$ ;  $^{137}\text{Cs} - 0,036$  %. При этом на 1 Ки/км<sup>2</sup> по  $^{90}\text{Sr}$  приходилось 4,6 Ки/км<sup>2</sup>  $^{95}\text{Zr} + ^{95}\text{Nb}$ , 12,2 Ки/км<sup>2</sup>  $^{144}\text{Ce}$  и около 0,7 Ки/км<sup>2</sup>  $^{106}\text{Ru}$ .

Были рассчитаны эквивалентные дозы облучения на отдельные органы (красный костный мозг, желудок, тонкий кишечник), нормированные на плотность поверхностного загрязнения 1 Ки/км<sup>2</sup> по  $^{90}\text{Sr}$ .

Аналогичного масштаба исследования проведены и по другим блокам. В результате имеется полная картина как экологического, так и социального состояния территории и всех приоритетных факторов в каждом блоке, формирующих здоровье населения.

Для обработки, представления и анализа информации использовались геоинформационные системы ArcView и MapInfo и их модификации. На электронную топографическую основу накладывались информационные слои вместе с соответствующим программным обеспечением, позволяющим проводить интерполяционные и прогностические процедуры для каждого фактора и по их комбинациям, а также зонирование территории как по отдельным факторам, так и по их совокупности.

Например, определен суммарный показатель загрязнения атмосферы, или индекс загрязнения атмосферы, г. Каменска-Уральского с годичным периодом усреднения. Аналогичные карты имеются на каждый день, неделю, месяц, сезон. В показа-

тель входят 55 веществ, наиболее значимых для г. Каменска-Уральского: тяжелые металлы и их оксиды, органические и неорганические соединения. Для каждого вещества и соответствующих групп суммаций имеются карты уровней загрязнения с разными периодами усреднения и карты максимальных разовых концентраций пыли на территории города.

Здоровье населения. Второй иерархический уровень параметров здоровья населения составляют рождаемость, смертность и заболеваемость. При этом в блок параметров рождаемости входят, кроме того, патология беременности и родов и состояние здоровья новорожденных. В параметрах смертности учитываются перинатальная и младенческая смертность, смертность по полу, возрасту, другим причинам и онкологическая смертность. Заболеваемость подразделяется на первичную и общую по классам, в том числе отдельно исследуется онкологическая заболеваемость детей, подростков и взрослых.

Поскольку глобальная цель – выявление роли радиационного фактора в формировании здоровья населения, то оценка параметров здоровья производилась в рамках традиционного подхода: случай – контроль или копия – пара, где в качестве основной территории была радиоактивно-загрязненная, а контрольной – территория, максимально близкая по экологическим, климатическим, социальным и другим параметрам, но отличающаяся по интенсивности радиоактивного загрязнения.

Исходя из этих требований, например для Каменского района, был выбран Артинский район, расположенный также на юге Свердловской области и не попавший под радиоактивное загрязнение. Для обоих районов было проведено сравнение всех параметров здоровья. В качестве примера в табл. 1 приведены показатели первичной заболеваемости взрослого населения.

Таблица 1. Основные показатели первичной заболеваемости взрослого населения Каменского и Артинского районов Свердловской области, на 1 000 человек

Класс болезней	Каменский р-н	Артинский р-н (контроль)
Болезни органов дыхания	254	157
Болезни нервной системы и органов чувств	117	64
Болезни крови	0,9	0,7
Болезни системы кровообращения, в т. ч. гипертоническая болезнь	103 21	44 12
Заболеваемость органов пищеварения, в т. ч. язвенная болезнь желудка и 12-перстной кишки	45 10	37 6
Болезни мочеполовой системы	44	29
Осложнения беременности и родов	10	6
Болезни костно-мышечной системы	57	36
Врожденные аномалии	0,5	0,1

Для дальнейших исследований на третьем иерархическом уровне и, соответственно, последующей оценки роли радиационного фактора выбирались болезни или параметры, различие в показателях которых было статистически достоверно.

К третьему иерархическому уровню оценки параметров здоровья относится донозологическая диагностика: компьютерная диагностика, биохимический анализ жидких сред организма и определение содержания вредных веществ в биосубстратах, физиологические параметры, иммунологический статус, комиссионное обследование.

В большинстве донозологических исследований использовался созданный в 1991 г. в Институте промышленной экологии УрО РАН программно-аппаратный комплекс ВИТА-МАРС, в основе которого заложены методики, разработанные Р. М. Баевским и сотр. (Институт медико-биологических проблем) и А. П. Берсеновой и сотр. (Московский областной научно-исследовательский клинический институт им. М. Ф. Владимирского).

Комплекс состоит из компьютера, стандартного одноканального кардиографа, интерфейсной платы связи компьютера с кардиографом и собственно программного обеспечения (экспертная система). Комплекс обеспечивает сбор, хранение и первичную обработку данных.

Методика ВИТА представляет собой согласованную систему сбора и частичного анализа данных, включающую анкетный опрос пациента (паспортные данные, анамнез, жалобы, условия жизни), антропометрические измерения, кардиологические исследования, данные осмотра средним медицинским персоналом. Анкетный опрос может проводиться как в диалоговом режиме, так и посредством опросных бланков, заранее раздаваемых пациентам.

На основании первичных данных формируются показатели гемодинамики, профили вероятной патологии, факторы риска и т. д., а также обобщающий параметр, называемый индексом функциональных изменений (ИФИ).

Методика МАРС (математический анализ ритма сердца) может быть использована как независимый инструмент, так и в составе методики ВИТА. В последнем случае результаты обработки по методике МАРС учитываются ВИТОЙ как дополнительные параметры.

Методика МАРС основана на математическом анализе последовательности кардиоинтервалов (100 временных промежутков между отдельными сердечными сокращениями, регистрируемых кардиографом). Авторы методики МАРС связывают вариабельность кардиоинтервалов с различными видами регуляции внутри организма. Результатом расчетов являются показатели функционирования отдельных контуров регуляторной системы организма. На их основе выводится показатель состояния регуляторной системы в целом. Методика первоначально разрабатывалась для оценки состояния регуляции организма космонавтов во время полета, поэтому наиболее удобна для оценки состояния взрослых практически здоровых людей.

Программно-аппаратный комплекс ВИТА-МАРС применялся в ИПЭ УрО РАН для определения состояния здоровья населения во многих экологических исследованиях, в том числе на территории Восточно-Уральского радиоактивного следа, бассейне р. Течи и на контрольных территориях. Методики использовались как самостоятельно, так и в комплексе с другими методами оценки здоровья (данные статистики, оценки биовозраста, осмотры специализированными медицинскими бригадами, специфические лабораторные методы и т. д.).

### **Оценка вклада радиационного и других факторов в формирование здоровья населения**

После проведения комплекса исследований параметров или факторов третьего иерархического уровня мы имеем полную структурированную информацию о

состоянии и динамике изменения радиационной и полной экологической обстановки, социальной и демографической ситуации, т. е. полный набор ранжированных по степени значимости параметров, влияющих на здоровье населения, знаем, какие параметры здоровья на радиоактивно-загрязненной территории достоверно отличаются от их значений на контрольных территориях. Одним словом, имеем всю необходимую информацию для установления корреляции между параметрами здоровья и факторами, формирующими его. Эта связь устанавливается при помощи метода распознавания образов. Метод достаточно мощный и универсальный, но, естественно, имеющий свои ограничения, которые нужно знать, чтобы его эффективно применять.

Пожалуй, основное ограничение состоит в том, что «заметить» влияние на здоровье (глобальная цель) можно только тех параметров или факторов, которые отличаются для исследуемой и контрольной групп населения. И чем сильнее различие, тем достовернее результаты.

Рассмотрим методологию распознавания образов на следующем примере.

1. Обследуются дети 4–7 лет, проживающие в Ленинском поселке г. Каменска-Уральского. Этот поселок вследствие аварии 1957 г. был загрязнен радионуклидами наиболее сильно. Подчеркнем, что все дети живут в одном месте.

2. Формируются две группы детей: эпизодически болеющие (ЭБ) и часто болеющие (ЧБ). Принципы формирования групп четко определены: если ребенок 7 лет болел менее 11 раз за всю жизнь, он помещается в группу ЭБ, а если более 20 раз, то в группу ЧБ. В каждой группе примерно по 300 человек. Можно сформировать три и более групп, все принципы сохраняются.

3. Вопрос: влияет ли на здоровье этих детей то обстоятельство, что их предки проживали или не проживали на радиоактивно-загрязненной территории в первые годы после аварии?

4. Определяется набор признаков, влияющих на здоровье детей:

- наличие предков, проживавших на радиоактивно-загрязненной территории в первые годы после аварии (11 признаков, характеризующих факты проживания на радиоактивной территории отца ребенка, матери ребенка, бабушки и бабушки по отцовской и материнской линиям и т. д.);
- уровень общетехногенного загрязнения территории в месте жительства ребенка и в месте расположения детского сада, который посещает ребенок (59 признаков, характеризующих загрязнение атмосферы, почвы и снега);
- уровень и качество жизни и демографическая характеристика семьи (31 признак);
- образ и условия воспитания в детском саду (19 признаков);
- индивидуальные особенности роста и развития ребенка (39 признаков).

Всего 159 признаков. Кстати, уровень радиоактивного загрязнения не включен в число признаков, поскольку на всей территории он примерно одинаков. Таким образом, в  $n$ -мерном пространстве ( $n = 159$ ) имеется облако точек (~600), каждая из которых характеризует ребенка и среду, в которой он находится.

5. Суть метода распознавания образов заключается в том, чтобы из всего набора признаков выбрать такие, которые позволяют разделить «облако» на две части. При этом определяется также эффективность, или информативность, признака по его роли в этом процессе, т. е. ранговое место.

6. Был найден высокоинформативный набор из 20 признаков, обеспечивающих надежную классификацию (разделение детей на ЭБ и ЧБ). По общим принципам это именно те факторы, которые влияют на заболеваемость детей.

7. Признаки, характеризующие наличие облученных предков, заняли места: 10-е – наличие лучевой экспозиции у бабушки по линии матери; 12-е – наличие лучевой экспозиции у дедушки по линии матери; 13-е – общее количество облученных предков по линии матери и т. д.

8. На первых ранговых местах по влиянию на заболеваемость детей оказались признаки, характеризующие условия пребывания и организацию воспитательно-образовательного процесса в детском саду: тяжесть социальной адаптации при поступлении в детский сад, инсоляционный режим в основных помещениях детского сада, проектная вместимость детского сада и др.

9. Признаки, характеризующие химическое загрязнение местности, вообще не вошли в информативный набор, поскольку, как и в случае радиоактивного загрязнения, общетехногенное загрязнение территории, где расположены детские сады, хотя и сравнительно высоко, примерно везде одинаково. Главное, что поставленная цель – определить, является ли радиационный фактор (облучение предков) доминирующим или нет в формировании здоровья, – достигнута.

### **Оценка ущерба, нанесенного территории радиоактивным загрязнением**

Применение метода системного анализа привело к довольно неожиданным результатам – суммарный ущерб оказался на порядок выше по сравнению с традиционным, «экономическим» подходом. В данном случае глобальная цель – оценка ущерба среди основных факторов первого уровня – экологические, социально-экономические и факторы здоровья.

На втором уровне из экологических факторов рассматривается только радиационный, общетехногенное загрязнение не учитывается. К социально-экономическим факторам добавляется блок непосредственного экономического ущерба, связанного с затратами на переселение жителей, реабилитацию загрязненных земель, с выведением из оборота земельных угодий и соответствующим недобором сельскохозяйственной продукции и со сносом социально-хозяйственных объектов. К ним же относятся социальные выплаты пострадавшему населению.

К демографическим факторам, существенным для оценки состояния здоровья, таким как численность населения школьного и дошкольного возраста, распределение населения в трудоспособном возрасте по возрастным группам, средний возраст и возрастной состав населения, образовательный уровень взрослого населения, добавляются: недожитие в трудоспособном возрасте и нанесенный ущерб, связанный с недопроизводством продукции; депопуляция населения, вызванная деградацией структурного распределения семей и как следствие уменьшением рождаемости; депопуляция за счет миграции. Все это ведет к уменьшению производства продукции на территории.

Ко всем факторам здоровья, названным выше, добавляются стохастические эффекты, обусловленные величиной накопленной дозы. Учет всех этих факторов и привел к нетривиальному результату.

### **Системный подход и экологические проблемы региона**

Необходимым условием корректной оценки состояния части сложной взаимосвязанной и взаимозависимой системы является учет влияния воздействия всех остальных ее частей на выбранную или учет всех факторов, определяющих состояние данной части системы. Игнорирование этого условия и попытка связать изменение состояния одной части системы с изменением другой части без учета осталь-



ных приводит к неверным выводам. Например, общий принцип: «чем хуже экологическая ситуация, тем хуже состояние здоровья населения» справедлив только при равных: уровне и качестве жизни, уровнях культуры, образования, медицинского обслуживания, развития экономики и т. д.

Даже для таких больших систем, как области и республики Уральского региона, не говоря уже о городах и районах, нельзя напрямую связывать состояние здоровья населения с экологической напряженностью.

В качестве характеристики экологической ситуации рассмотрим состояние атмосферы, поверхностных вод и количество накопленных токсичных отходов.

По количеству суммарных выбросов вредных веществ в атмосферу Уральский экономический регион (УЭР) на протяжении последних лет занимает первое место среди экономических регионов России. Вклад региона в загрязнение атмосферы стационарными источниками на протяжении 1990–1997 гг. составлял 23–26 %. Наибольший вклад в выбросы региона дают Свердловская и Челябинская области: 31 и 26 % соответственно.

Зонирование региона по степени загрязненности поверхностных вод на региональном уровне в первом приближении можно провести на основе объема сбрасываемых загрязненных вод и их суммарного приведенного индекса опасности.

Сброс загрязненных сточных вод на территории УЭР составляет 10,2 % от общего по России. На долю Свердловской области приходится 31 % общего объема загрязненных сточных вод УЭР, доля Челябинской области составляет 25, Башкортостана – 20, Пермской области – 18 %. Динамика выбросов в атмосферу и сброса загрязненных сточных вод в водные объекты по республикам и областям УЭР показана в табл. 2. Там же представлены значения индекса опасности сбросов.

Таблица 2. Показатели воздействия на окружающую среду в Уральском экономическом регионе

Область, республика	Выброс загрязняющих веществ в атмосферу от стационарных источников, тыс. т		Плотность эмиссии, т/(м <sup>2</sup> /год)	Сброс загрязненных сточных вод, млн м <sup>3</sup>		Приведенный индекс опасности сбросов	Колич. накопленных отходов, т/км <sup>2</sup>
	1991 г.	1997 г.		1997 г.	1991 г.		
Башкортостан	854,0	645,5	5,4	578,2	486,8	0,65	506
Курганская	174,0	113,9	1,6	25,5	19,7	0,44	1,3
Оренбургская	912,0	515,3	4,4	147,7	78,7	0,89	683
Пермская	874,0	570,7	4,0	610,8	379,5	16,9	9,3
Свердловская	2 439,1	1 361,4	7,2	835,2	827,3	1,23	3 360
Удмуртия	254,0	207,1	4,8	69,1	35,8	2,80	59
Челябинская	2061,0	1038,4	10,8	662,0	777,0	3,73	673
Российская Федерация			1,0				

$$J = \sum(K_i/\text{ПДК}_i)/S,$$

где  $K_i$  – ежегодный сброс  $i$ -го поллютанта на территории области или республики, т/год;  
 $\text{ПДК}_i$  – предельно допустимая концентрация  $i$ -го поллютанта в воде, мг/л;  
 $S$  – средний многолетний сток на территории, м<sup>3</sup>/год.

В качестве параметра для зонирования региона по количеству накопленных токсичных отходов можно использовать в первом приближении количество отходов, приходящихся на единицу площади рассматриваемой территории (см. табл. 2).

По данным Госкомстата России на 01.01.97, на предприятиях Уральского региона размещено около 400 млн т токсичных отходов всех классов опасности, из которых почти 40 % – на территории Свердловской области. Особую тревогу вызывают токсичные отходы первого класса опасности (отходы гальванических производств, отходы, содержащие ртуть, хлорорганику, хром и др.) Таких отходов ежегодно на предприятиях УЭР образуется 65,5 % от общего их количества по Российской Федерации.

Не меньшую опасность представляет накопление на промплощадке ПО «Маяк» в Челябинской области радиоактивных отходов. С учетом накопленных радиоактивных отходов Челябинская область выходит на первое место по степени неблагоприятия, связанной с накоплением токсичных отходов.

На основании значений критериев, рассмотренных выше, различные экспертные оценки приводят к следующему ранжированию территорий по степени экологического неблагоприятия или экологической напряженности (табл. 3).

Таблица 3. Ранжирование территорий Уральского региона на основе экспертных оценок экологического неблагоприятия и ухудшения здоровья населения

Территория	Экологическая ситуация	Здоровье населения
Республика Башкортостан	2 – 3	7
Удмуртская республика	6	3 – 4 – 5
Курганская область	7	2
Оренбургская область	4	3 – 4 – 5
Пермская область	5	6
Свердловская область	2 – 3	3 – 4 – 5
Челябинская область	1	1

Причем при самых различных вариантах бесспорный лидер – Челябинская область, а аутсайдер – Курганская область. Аналогичное зонирование региона по состоянию здоровья населения, включающее первичную заболеваемость по разным видам болезней, новообразования, врожденные аномалии, болезни органов дыхания, пищеварения, эндокринной и костно-мышечной систем, болезни кожи, грипп, ОРЗ и т. д. (всего 13 показателей с последующей экспертной оценкой), привело к другой последовательности (см. табл. 3). Можно сделать вывод, что для Челябинской области ведущим фактором, определяющим состояние здоровья населения, является экологическая обстановка. Для Свердловской, Пермской и Орен-

бургской областей экологический фактор один из определяющих. Для Курганской области, Башкортостана он является фактором второго порядка.

Таким образом, экологические принципы, справедливые в своей среде, но перенесенные на более высокий уровень обобщения, оказываются неэффективны. В этом причина неудач федеральных, региональных и локальных экологических программ.

### **Статус зоны чрезвычайной экологической ситуации – системный подход**

Институт осуществляет комплексное междисциплинарное исследование и решение проблем экологически неблагоприятных территорий. Разработанный в Институте единый подход, основанный на методологии системного анализа, применяется для комплексного исследования широкого круга экологических, экономических, социальных, медицинских и технологических проблем территорий различного уровня с последующей разработкой и сопровождением программ выхода из кризиса.

Иерархическая схема комплексной оценки состояния территории и установления приоритетов включает в себя:

- Первый блок – состояние среды обитания
- Второй блок – антропогенные источники воздействия на окружающую среду.
- Третий блок – социально-экономическая среда территории.
- Четвертый блок – состояние здоровья населения.
- Пятый блок – комплексная оценка состояния территории – системный анализ.

При исследованиях первого блока по оценке среды обитания рассматривается состояние компонентов окружающей среды:

- атмосферного воздуха, поверхностных и подземных вод, почвы, а также
- продуктов питания, состояние растительности, радиоактивное загрязнение.

Качество среды обитания оценивается системой санитарно-гигиенических, общеэкологических и других требований. Задача исследований этого блока – получить подробную картину распределения и концентрации загрязняющих веществ на территории.

Исследования второго блока включают характеристику антропогенных источников воздействия:

- перечень технологий, применяемых на территории промышленной зоны;
- виды и объемы загрязнений, поступающих в окружающую среду, с ранжированием их по степени опасности;
- анализ эколого-экономической эффективности деятельности предприятий.

Цель проводимого анализа – выделить основные объекты, оказывающие приоритетное воздействие на окружающую среду, и определить пути его уменьшения.

Работы третьего блока по оценке социально-экономической среды территории включают в себя оценки:

- социально-демографической ситуации;
- уровня и качества жизни;
- финансово-экономического положения территории;
- эколого-градостроительную оценку.

Целью исследований этого блока является характеристика социальной инфраструктуры и анализ финансового потенциала территории для проведения природоохранных мероприятий.

Исследования четвертого блока по оценке состояния здоровья населения (рождаемости, смертности, заболеваемости, распределению ее по половозрастным признакам и классам болезней и т. д.) проводятся на основе:

- материалов медицинской статистики;
- данных скрининг-обследования населения;
- результатов углубленных лабораторно-клинических обследований групп населения, выбранных по семейному принципу.

В результате исследований этого блока дается оценка состояния здоровья населения в сравнении с контролем и тенденция его изменения. По результатам углубленных исследований групп населения возможно выявление основных факторов риска потери здоровья.

Пятый блок представляет собой системный анализ всей информации с использованием методов математической статистики, где главенствующую роль играет метод распознавания образов. В результате определяются приоритетные экологические проблемы в плане их влияния на здоровье населения, экономику, социально-демографическую ситуацию. Как правило, приоритеты не совпадают.

На базе проведенного анализа формируется программа социально-экономического развития в условиях неблагоприятной экологической обстановки.

Программа имеет следующие основные цели:

- стабилизацию и доведение до нормативов показателей состояния окружающей среды;
- реабилитацию состояния здоровья населения;
- разработку системы социальных компенсаций населению за проживание в условиях экологического кризиса.

На основании результатов анализа определяются экологический статус территории и степень его соответствия статусу зоны чрезвычайной экологической ситуации или зоны экологического бедствия.

В Институте по заданию Минприроды проведена апробация критериев для определения экологического статуса территории и на основе этих критериев впервые в стране обоснован статус города как зоны чрезвычайной экологической ситуации (г. Каменск-Уральский) и разработана Федеральная целевая программа первоочередных мер по нормализации обстановки, утвержденная правительством РФ. Аналогичная работа выполнена для Орска, Тулы, Липецка.

Статус присваивается, если есть соответствующие отклонения в состоянии здоровья и соответствующие превышения в загрязнении хотя бы двух из трех сред, например, почвы и атмосферы.

Основная цель программ – реабилитация здоровья населения. Но разрабатываются они на экологических принципах, т. е. на нейтрализации в первую очередь приоритетных загрязнителей – загрязнителей, имеющих наибольшие концентрации в относительных единицах и наибольшие ареалы загрязнения. Но совершенно необязательно, что медико-экологические приоритеты будут такие же.

Здесь трудно что-либо изменить, и мы вынуждены разрабатывать программы по правилам Минприроды, хотя и видим их слабые стороны. Правда, для г. Каменска-Уральского удалось обосновать статус, не включая радиационный фактор в число определяющих.

Кстати, впервые в стране работа такого масштаба и комплексности проводилась в 1990–1992 гг. для Первоуральско-Ревдинского промышленного района. Именно здесь апробировались и разрабатывались идеи методологии комплексного анализа.

## **Другие основные проблемы, над которыми работает Институт**

### **Радиационные риски**

Оценка радиационного воздействия на население региона от различных источников, а также обусловленных ими радиационных рисков и является основной задачей радиационной лаборатории.

Оценка облучения населения Свердловской области за счет радона и его дочерних продуктов потребовала проведения измерений как объемной активности  $^{222}\text{Rn}$  либо его дочерних продуктов распада, так и измерений эквивалентной равновесной объемной активности (ЭРОА)  $^{220}\text{Rn}$  (торона). Обследования уровней накопления радона были проведены в 10 районах и 5 городах Свердловской области. В программу обследований вошли 80 населенных пунктов. В тех населенных пунктах, в которых проводились измерения, проживает около 2 млн человек, или 43 % населения Свердловской области. Обследовались сельские дома, городские квартиры и детские учреждения.

Распределение значений ЭРОА радона и торона подчиняется логнормальному закону. Известный вид распределения и его параметры позволяют прогнозировать долю помещений, в которых ЭРОА радона превысит определенные референсные уровни. В качестве референсных в данном представлении выбраны величины 100 и 200 Бк/м<sup>3</sup> ЭРОА радона, 10 и 20 Бк/м<sup>3</sup> ЭРОА торона.

Уровни накопления радона в помещениях, наряду с типом здания, определяются такими важными факторами, как георадиохимические и геолого-геофизические характеристики территории. Поэтому полученные данные по величинам ЭРОА радона в воздухе помещений были разделены в соответствии с комплексом таких признаков.

### **Оценка природных и техногенных рисков для населения, экономики и окружающей среды Свердловской области**

В результате комплексной оценки природных и техногенных рисков для населения, экономики и окружающей среды Свердловской области на основании информации о существующих в области источниках опасности природного характера, опасных видах промышленной деятельности, угрозах экологической и экономической безопасности количественно оценены различные составляющие природного, техногенного и экологического риска. Количественно оценен индивидуальный, коллективный и экономический риск, обусловленный следующими видами опасности: опасные гидрологические ситуации, опасные метеорологические ситуации, лесные пожары, инфекционные заболевания, аварии на химически опасных объектах, аварии на взрыво- и пожароопасных объектах, аварии на магистральных трубопроводах, аварии на транспорте (железнодорожный, автомобильный, авиационный), рак в результате химического загрязнения окружающей среды (атмосферного воздуха и питьевой воды), рак в результате облучения продуктами распада радона, ущерб окружающей среде в результате чрезвычайных ситуаций. Осуществлено ранжирование рисков различной природы и проведено районирование территории области

по уровням природного и техногенного риска. Для Свердловской области подобная комплексная работа выполнена впервые.

В Институте сформировалось и получило развитие новое для УрО РАН направление, связанное с комплексным изучением последствий деятельности предприятий атомной промышленности и энергетики. В Институте успешно выполнена работа по оценке воздействия на окружающую среду проекта строительства III очереди Белоярской АЭС с реактором БН-800. Это был первый в стране проект строительства нового блока АЭС, по которому получено положительное заключение Минэкологии. Институт принимал участие в работах по экологическому обоснованию возможности строительства Южно-Уральской АЭС. Выполнена оценка воздействия на окружающую среду ряда тепловых станций, в частности, Чайковской.

### **Оценка воздействия на окружающую среду и экологическое прогнозирование**

Решением этих проблем в основном занимается лаборатория экологического прогнозирования. Основное направление ее деятельности – разработка научно-методического обеспечения экологического мониторинга, оценки и управления воздействием на окружающую среду объектов повышенного экологического риска, в частности, предприятий горно-металлургического комплекса.

В лаборатории разработана комплексная методика, позволяющая успешно решать вопросы обеспечения экологической безопасности хозяйственной деятельности на всех этапах инвестиционного процесса.

Методика базируется на всестороннем анализе факторов воздействия горного предприятия на окружающую среду и результатов комплексных натуральных экспериментов и включает в себя три взаимосвязанных блока исследований:

- оценку воздействия на окружающую среду на стадии разработки предпроектной и проектной документации на строительство или реконструкцию предприятия;
- разработку и реализацию системы экологического мониторинга зоны влияния предприятия;
- разработку системы мероприятий по уменьшению первичных и отдаленных экологических и социально-экономических последствий деятельности предприятий.

Основные методические подходы разработаны и опробованы при выполнении экологического обоснования по проекту вскрытия и разработки Сафьяновского медно-рудного месторождения в Режевском районе Свердловской области в 1993 – 1994 гг.

Впервые в Уральском регионе выполнена оценка воздействия на окружающую среду по проекту до начала его реализации, что позволило на стадии проектирования предусмотреть необходимые меры и технические решения для уменьшения возможных отрицательных экологических и социально-экологических последствий.

Разработанная методика была реализована и на других крупных горно-металлургических предприятиях Урала, в частности, Высокогорском горно-обогатительном комбинате (г. Нижний Тагил), Гайском горно-обогатительном комбинате (г. Гай Оренбургской области) и др.

## **Заключение**

Урал продолжает оставаться одним из самых экологически неблагополучных регионов России. Решение его экологических проблем только за счет проведения мероприятий компенсационного типа на сегодня невозможно ни технологически, ни экономически. Решать проблему необходимо комплексно, согласуясь с экономическими, социальными и природно-экологическими реалиями одновременно. Только на основе системного анализа всей совокупности действующих факторов самой различной природы – от физико-химической до социальной и комплексной оценки последствий – возможно определение зависимости между масштабами и уровнями загрязнения окружающей среды и масштабами последствий для территории (здоровья населения, социальной среды, экономики и т. д.) и установление экологических приоритетов с последующей разработкой программы реабилитационных мер.