Федеральное государственное бюджетное учреждение науки
Институт промышленной экологии
Уральского отделения Российской академии наук

На правах рукописи

Буторова Анастасия Сергеевна

ВЫЧИСЛИТЕЛЬНОЕ МОДЕЛИРОВАНИЕ И РАЗВИТИЕ ПОДХОДА LAND USE В ЗАДАЧЕ ПОСТРОЕНИЯ ПРОСТРАНСТВЕННОГО РАСПРЕДЕЛЕНИЯ ПРИМЕСЕЙ В КОМПОНЕНТАХ ОКРУЖАЮЩЕЙ СРЕДЫ

1.2.2 – Математическое моделирование, численные методы и комплексы программ

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата физико-математических наук

Научный руководитель
к. ф.-м. н., доцент
Сергеев Александр Петрович

Екатеринбург – 2026

Содержание

Список основных обозначений

Обозначение	Описание
Ω	исследуемая территория
$\Omega' \subset \Omega$	область источников примеси
Ω'_m, $m = 1, \ldots, M$	m-й сегмент области источников, M – число сегментов
Θ	пространство допустимых значений параметров модели
$L_2(\Omega)$	пространство квадратично интегрируемых функций на Ω
$\boldsymbol{s} = (x, y)$	произвольная точка исследуемой территории
$\boldsymbol{s_i} = (x_i, y_i)$, $i = 1, \ldots, n$	i-я точка наблюдения, n – число точек наблюдений
$\boldsymbol{s'} = (x', y')$	точка, принадлежащая области источников
$\boldsymbol{s'_m}$	центроид m-го сегмента источников
$d(\boldsymbol{s}, \boldsymbol{s'}) = \|\boldsymbol{s} - \boldsymbol{s'}\|$	расстояние между точками $\boldsymbol{s}$ и $\boldsymbol{s'}$, $\|\cdot\|$ – евклидова метрика
$\boldsymbol{\theta}$	вектор параметров модели
$\widehat{\boldsymbol{\theta}}$	оценка вектора параметров модели
$f(\boldsymbol{a}, \boldsymbol{\theta})$	функция вклада источника
$\boldsymbol{a(s, s')}$	вектор аргументов функции вклада, зависящих от положений точек $\boldsymbol{s}$ и положений источников $\boldsymbol{s'}$
$C(\boldsymbol{s})$	содержание примеси в точке $\boldsymbol{s}$
$C^{obs}(\boldsymbol{s_i})$	наблюдаемое содержание примеси в точке $\boldsymbol{s_i}$
$\hat{C}(\boldsymbol{s}, \boldsymbol{\theta})$	модельная оценка содержания примеси
$\varphi(\boldsymbol{s})$	глобальный пространственный тренд
$Loss(\boldsymbol{\theta})$	функция потерь
$Buffer(\boldsymbol{s_i}, r_k)$	круговая окрестность точки $\boldsymbol{s_i}$ радиуса r_k
$Ring(\boldsymbol{s_i}, r_{k-1}, r_k)$	кольцевая окрестность точки $\boldsymbol{s_i}$ (r_{k-1}, r_k)
r_k, $k = 1, \ldots, K$	внешний радиус k-й окрестности, K – число окрестностей
$\mathbf{X}$	матрица пространственных переменных

$v_{ij}, j = 1, \ldots, p$	значение j-го предиктора в i-й точке наблюдения, p – число предикторов модели
$\mathbf{X}^{\mathrm{T}}\mathbf{X}$	матрица Грама
VIF_j	фактор инфляции дисперсии j-го предиктора
$\kappa(\mathbf{X})$	число обусловленности матрицы $\mathbf{X}$
RMSE	среднеквадратическая ошибка
MAE	средняя абсолютная ошибка
R^2	коэффициент детерминации
Corr	коэффициент корреляции Пирсона
p-value	*p-value* пермутационного теста
Train	обучающее подмножество модели
Test	тестовое подмножество модели
$R(\boldsymbol{s}_i)$	репрезентативность точки наблюдения $\boldsymbol{s}_i$
$O(\cdot)$	асимптотическая вычислительная сложность

Введение

Актуальность темы исследования. При оценке состояния компонентов окружающей среды широко используются агрегированные показатели, такие как средние по территории значения содержаний примесей [1, 2]. Использование исключительно средних по территории значений может приводить к искаженной оценке экологической обстановки: при равных средних значениях пространственная структура распределения примесей может быть настолько неравномерной, что по воздействию на здоровье населения такие территории оказываются совершенно различными [3]. Учет пространственной структуры таких распределений может сделать принятие решений в сфере экологического мониторинга и управления более обоснованным.

Задача построения пространственного распределения примесей в компонентах окружающей среды предполагает аппроксимацию непрерывного поля по результатам измерений в конечном множестве точек наблюдений. Современные подходы к моделированию позволяют строить поля с высоким пространственным разрешением и учитывать локальные неоднородности, однако степень пространственной подробности таких полей определяется плотностью сети наблюдений и репрезентативностью данных измерений. В задачах экологического мониторинга эти условия часто не выполняются. Сбор данных о состоянии окружающей среды может быть затруднен из-за высокой стоимости необходимого оборудования, труднодоступности территорий и сезонных ограничений наблюдений. Особенно это проявляется в северных регионах, где проведение систематических измерений дополнительно сопряжено с организационными и финансовыми трудностями [4–7]. Указанные обстоятельства часто не позволяют получить достаточный объем репрезентативных данных. Модели, обученные на таких наборах данных, оказываются неустойчивыми, а прогнозы – ненадежными.

При построении моделей пространственного распределения примесей, наряду с данными наблюдений, может использоваться дополнительная информация о территории – структура землепользования. Данные о

землепользовании включаются в модель в качестве пространственных характеристик территории, предположительно связанных с содержанием примеси в точках наблюдений. Такой подход компенсирует недостаток данных наблюдений и повышает пространственное разрешение моделируемых полей.

Существующие подходы к построению полей пространственного распределения примесей могут быть представлены двумя классами моделей: физическими моделями и моделями аппроксимации (в частном случае – интерполяции). Физические модели решают уравнения переноса и рассеяния примесей с учетом метеоусловий, характеристик источников примесей и свойств подстилающей поверхности. Эти модели физически обоснованы и хорошо интерпретируемы, но требуют большого объема входных данных, чаще всего получаемых с помощью натурных измерений, точной калибровки измерительных приборов и значительных вычислительных ресурсов. Модели аппроксимации опираются на измерения в точках наблюдений и предполагают наличие пространственной зависимости между измерениями. Современные примеры таких моделей – модели машинного обучения и, в частности, нейронные сети, – позволяют достигать высокой точности аппроксимации, но, как правило, чувствительны к плотности и пространственной конфигурации наблюдений: при разреженных наборах данных такие модели могут терять устойчивость и прогностическую точность.

Отдельное место среди аппроксимационных подходов занимает Land Use, который использует данные о землепользовании для построения предикторов. Подход Land Use был разработан и впервые применен в 1990-х годах и получил широкое распространение в задачах оценки качества атмосферного воздуха [8]. Подход Land Use учитывает информацию о пространственном взаиморасположении точек наблюдений и потенциальных источников примесей и других характеристик, так или иначе влияющих на процессы пространственного распределения примесей. Идея Land Use подхода состоит в том, что на основе доступных географических данных формируется набор пространственно информированных предикторов. Эти предикторы затем используются в модели для

построения поля пространственного распределения примеси. В классической реализации Land Use в качестве модели выступает линейная регрессия, в связи с чем часто используется термин Land Use Regression [9]. В настоящей работе подход Land Use рассматривается в более общем смысле – как самостоятельный этап построения предикторов на основе доступных географических данных, после которого может применяться любая модель аппроксимации – от линейной регрессии до нейронной сети.

К основным преимуществам подхода Land Use относятся использование географически информированных предикторов, более высокая продуктивность и возможность получения полей примесей с высоким пространственным разрешением при малом количестве наблюдений, что особенно ценно для гетерогенных сред, характеризующихся выраженной пространственной неоднородностью [10–18]. Вместе с тем классическая реализация Land Use Regression имеет ряд ограничений. Пространственные переменные, как правило, строятся на основе пересечения так называемых «буферных зон» (buffer zones/buffers) – концентрических окрестностей точек наблюдений (представляющих собой круги на проекции территории обычно с евклидовой метрикой) – с предполагаемыми источниками примеси. Такой подход приводит к многократному учету одних и тех же источников [19]. Это, в свою очередь, приводит к мультиколлинеарности предикторов, неустойчивости оценок коэффициентов регрессии и затрудняет интерпретацию расстояния, на котором источник оказывает влияние на содержание примеси в точке наблюдения. Указанные ограничения в настоящей работе определяют направления развития подхода Land Use и разработки новых моделей построения пространственного распределения примесей.

Цель и задачи диссертационной работы. Цель настоящей работы – развитие подхода Land Use, исследование и разработка математических моделей, соответствующих им численных методов и реализующих их комплексов программ для моделирования пространственного распределения примесей в компонентах

окружающей среды, обеспечивающих высокое пространственное разрешение получаемых полей примесей при ограниченных объемах данных наблюдений.

Для достижения поставленной цели были сформулированы следующие задачи:

1. Анализ существующих методов и подходов к построению математических моделей пространственного распределения примесей в компонентах окружающей среды, их ограничений при работе с малыми и разреженными наборами данных, обоснование необходимости использования подхода Land Use в настоящей работе.

2. Разработка Land Use подхода к построению пространственных переменных, снижающего мультиколлинеарность между предикторами модели, реализация его численных методов и анализ численной устойчивости моделей, пространственные переменные для которых построены с применением этого подхода.

3. Разработка математических моделей Land Use, в том числе физически мотивированной модели, и соответствующих им численных методов, снижающих ошибку аппроксимации пространственного распределения примесей в условиях ограниченных данных наблюдений, и реализация этих моделей и методов в виде программного комплекса.

4. Разработка метода оценки репрезентативности точек наблюдений и его применение для анализа информативности различных конфигураций точек наблюдений в задаче построения пространственного распределения примеси в условиях ограниченных данных наблюдений.

5. Тестирование разработанных моделей на синтетических данных и апробация на данных измерений примесей в различных компонентах окружающей среды, сравнение разработанных моделей с моделями на основе классического подхода к построению пространственных переменных Land Use на основе круговых окрестностей точек наблюдений (Land Use Buffers), оценка моделей с использованием традиционных метрик качества и пермутационного подхода.

Объект исследования: пространственные распределения примесей в компонентах окружающей среды.

Предмет исследования: численные методы и физически мотивированные модели пространственных распределений примесей в компонентах окружающей среды на основе данных наблюдений и данных о землепользовании.

Научная новизна диссертационного исследования:

1. Впервые предложен подход к построению пространственных переменных – предикторов для моделей Land Use, при котором пространственные переменные формируются на основе непересекающихся кольцевых окрестностей с центрами в точках наблюдений (Land Use Rings), и исследована численная устойчивость оценок коэффициентов регрессионной модели (п. 2 паспорта специальности 1.2.2 «Математическое моделирование, численные методы и комплексы программ»).

2. Впервые предложено развитие Land Use подхода – переход от пространственных переменных, формируемых на основе окрестностей точек наблюдений, к покрытию исследуемой территории непересекающимися множествами (элементами). Для каждого элемента покрытия определяется набор типов источников примеси и функции, описывающие его вклад в формирование пространственного распределения примеси на исследуемой территории. При этом функция вклада может быть задана любым подходящим способом (например, аналитическим выражением или нейросетевой моделью, в том числе физически мотивированной). В настоящей работе реализован частный случай предложенного подхода – разбиение территории на прямоугольные сегменты (Land Use Segmentation) (п. 1, 2 паспорта специальности 1.2.2 «Математическое моделирование, численные методы и комплексы программ»).

3. Впервые разработаны принципы применения пермутационного подхода к оценке качества моделей пространственного распределения примесей в компонентах окружающей среды. Пермутационный подход позволяет оценить вероятность получения наблюдаемого значения выбранной исследователем статистики различия/сходства и дополняет традиционные метрики качества моделей. Предложенный подход реализован в виде численного алгоритма и представляет собой непараметрический метод, не требующий предположений о виде распределений и допускающий использование любой подходящей статистики

(п. 5 паспорта специальности 1.2.2 «Математическое моделирование, численные методы и комплексы программ»).

4. Впервые предложен и реализован в виде численного алгоритма метод оценки индивидуальной и групповой репрезентативности точек наблюдений для задачи пространственного моделирования, основанный на выборе наилучших (в смысле минимизации/максимизации выбранной метрики различия/сходства между наблюдениями и прогнозом, соответственно) обучающих подмножеств (п. 4 паспорта специальности 1.2.2 «Математическое моделирование, численные методы и комплексы программ»).

5. Впервые разработана математическая модель пространственного распределения примесей на основе подхода Land Use Segmentation, в которой используется параметризованная функция вклада гауссовского типа (физически мотивированная модель) для построения пространственного распределения примесей при ограниченных данных наблюдений. Соответствующие численные методы реализованы в виде комплекса программ для вычислительного моделирования пространственного распределения примесей (п. 1, 3 паспорта специальности 1.2.2 «Математическое моделирование, численные методы и комплексы программ»).

Теоретическая и практическая значимость. Теоретическая значимость состоит в развитии методов построения математических моделей пространственного распределения примесей в компонентах окружающей среды на основе подхода Land Use. Предложенные подходы Land Use Rings и Land Use Segmentation развивают классическую идею построения пространственных переменных Land Use Buffers. Предложенный подход Land Use Rings снижает мультиколлинеарность между предикторами и повышает устойчивость моделей пространственного распределения примесей. Land Use Segmentation позволяет интегрировать в модель все доступные пространственные данные с различной степенью детализации и позволяет подбирать параметры функций вклада и/или веса нейросетевой модели таким образом, чтобы минимизировать расхождение между оценкой содержания примесей и данными измерений. Land Use

Segmentation повышает точность и снижает ошибку аппроксимации пространственного распределения примесей.

Практическую значимость составляют разработанные численные методы и алгоритмы и реализующий их комплекс программ для моделирования пространственного распределения примесей в компонентах окружающей среды. Разработанные алгоритмы позволяют повысить точность и снизить ошибку аппроксимации пространственного распределения примесей при ограниченных данных наблюдений. Полученные результаты могут быть использованы для решения прикладных задач экологического скрининга и мониторинга, в частности для построения карт распределения примесей, в том числе загрязняющих веществ, на территории.

Основные положения, выносимые на защиту:

1. Предложенный подход Land Use Rings к построению пространственных переменных на основе непересекающихся кольцевых окрестностей точек наблюдений снижает мультиколлинеарность предикторов модели, измеренную по фактору инфляции дисперсии, в среднем в 19,16 раза (бутстреп-оценка; 0,95 доверительный интервал: 18,79–19,84) и повышает устойчивость оценок коэффициентов регрессионной модели, измеренную по числу обусловленности матрицы $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, где $\mathbf{X}$ – матрица пространственных переменных, в 7,52 раза (бутстреп-оценка; 0,95 доверительный интервал: 7,35–7,67) по сравнению с классическим подходом Land Use Buffers к построению пространственных переменных на основе вложенных круговых окрестностей точек наблюдений в серии численных экспериментов.

2. Подход Land Use Segmentation, основанный на разбиении территории непересекающимися сегментами с заданными функциями вклада, повышает коэффициент корреляции между предсказанными и наблюдаемыми значениями на 0,3–2,3% и снижает ошибку аппроксимации пространственного распределения примеси *RMSE* на 22–32%, что свидетельствует о увеличении точности построения пространственных распределений примесей по сравнению с подходами на основе

окрестностей точек наблюдений Land Use Buffers и Land Use Rings в серии численных экспериментов.

3. В задаче построения пространственного распределения тяжелых металлов в верхнем слое почвы точки наблюдений обладают неравной репрезентативностью, оцениваемой как среднее значение *RMSE* моделей, обученных на подмножествах, содержащих эту точку. Наиболее репрезентативными являются точки, включение которых в обучающее подмножество обеспечивает минимальную ошибку *RMSE* на тестовом подмножестве.

Апробация результатов. Основные результаты диссертации докладывались на следующих конференциях:

1. 9th International Conference on New Trends of the Applications of Differential Equations in Sciences (NTADES 2022), Созополь, Болгария, 14–17 июня 2022;

2. International Conference of Numerical Analysis and Applied Mathematics: (ICNAAM 2022), Ираклион, Греция, 19–25 сентября 2022;

3. VII-th International Symposium «Topical Problems of Nonlinear Wave Physics» Москва–Санкт-Петербург, Россия, 7–13 сентября 2025;

4. Научно-практическая конференция «Биосферная совместимость атомной энергетики», Екатеринбург, Россия, 3–4 марта 2026;

5. 2026 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), Екатеринбург, Россия, 14–15 мая 2026.

Публикации. Материалы диссертации опубликованы в 18 работах, из них 5 статей, соответствующих требованиям ВАК как входящих в наукометрические базы данных Scopus и Web of Science, 1 статья в журнале перечня ВАК, 2 патента на изобретения и 1 свидетельство о государственной регистрации программы для ЭВМ.

Степень достоверности результатов проведенных исследований. Достоверность результатов исследования обусловлена корректным использованием положений и методов математического моделирования в задаче

построения пространственного распределения примесей, а также валидацией разработанных моделей с использованием данных наблюдений.

Соответствие паспорту научной специальности. Диссертация соответствует п. 1 «Разработка новых математических методов моделирования объектов и явлений», п. 2 «Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий», п. 3 «Реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента», п. 4 «Разработка новых математических методов и алгоритмов интерпретации натурного эксперимента на основе его математической модели» и п. 5 «Разработка новых математических методов и алгоритмов валидации математических моделей объектов на основе данных натурного эксперимента или на основе анализа математических моделей» паспорта специальности 1.2.2 – Математическое моделирование, численные методы и комплексы программ.

Личный вклад автора. Содержание диссертации и основные положения, выносимые на защиту, отражают личный вклад автора в опубликованные работы. Постановка задач, обсуждение полученных результатов и подготовка публикаций проводились совместно с научным руководителем и соавторами опубликованных работ. Все результаты, составляющие основное содержание диссертации, получены автором самостоятельно. Используемые в работе данные наблюдений получены соавторами публикаций.

Структура и объем диссертации. Диссертация состоит из введения, трех глав, заключения, библиографии и 4 приложений. Общий объем диссертации составляет 161 страницу, включая 34 рисунка и 17 таблиц. Библиография содержит 133 наименования на 15 страницах.

Глава 1. Подходы к построению пространственного распределения примесей в компонентах окружающей среды (обзор литературы)

Задача построения поля пространственного распределения примеси может быть определена в общем виде как задача восстановления неизвестной функции по ограниченному набору наблюдений [20–22]. В настоящей работе рассматривается геохимическое поле – область геологического пространства, в которой распределение содержания примеси $C(\boldsymbol{s}, t)$ представляется функцией пространственных координат $\boldsymbol{s} = (x, y)$ и, в общем случае, времени t.

В настоящей работе рассматривается стационарная постановка задачи, в которой временная зависимость учитывается неявно через интегральные характеристики накопления примеси в депонирующих средах. В рамках данной постановки задача построения поля формулируется как задача оценки функции $C(\boldsymbol{s})$, наилучшим образом согласующейся с наблюдаемыми значениями примеси в конечном наборе точек наблюдений:

$$\hat{C}(\boldsymbol{s}) = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (C(\boldsymbol{s_i}) - f(\boldsymbol{s_i}))^2, \qquad (1.1)$$

где $\boldsymbol{s_i} = (x_i, y_i)$, $i = 1, \dots, n$ – координаты точек наблюдений, n – число наблюдений, $C(\boldsymbol{s_i})$ – измеренные значения содержания примеси в точках наблюдений $\boldsymbol{s_i}$, f – аппроксимирующая функция, $\mathcal{F}$ – функциональный класс допустимых решений, $\hat{C}(\boldsymbol{s})$ – оценка пространственного распределения примеси, соответствующая решению задачи минимизации.

Для решения поставленной задачи применяются различные подходы, обзор которых приведен в настоящей главе.

1.1 Традиционные подходы

1.1.1 Физические модели массопереноса

Физические модели массопереноса описывают процессы переноса примесей от источников с учетом факторов, влияющих на конечное содержание примеси: химических превращений, депозиций, атмосферной турбулентности и

метеорологических условий [20–23]. Базовым уравнением, описывающим динамику содержания примеси, является уравнение адвекции-диффузии:

$$\frac{\partial C}{\partial t} + \mathbf{u} \cdot \nabla C = \nabla \cdot (D \nabla C) + S, \qquad (1.2)$$

где $C(\boldsymbol{s}, t)$ – содержание примеси, $\mathbf{u}$ – вектор скорости потока, D – тензор турбулентной диффузии, S – члены источников и стоков (эмиссия, химические реакции, осаждение и т. д.).

Таким образом физические модели используются для построения поля примеси при известных характеристиках источников и параметрах среды. Современная классификация физических моделей массопереноса основывается на нескольких критериях: масштаб применения модели (локальный, городской, региональный, глобальный), состав учитываемых факторов (физических, химических и биологических, таких как диффузия, конвекция, химические реакции, взаимодействие с поверхностями), тип численного описания (разнообразные аналитические, численные, гибридные методы) и подход к моделированию (детерминистический, стохастический, статистический или гибридный) [24].

На практике при нормировании выбросов и проведении экологических экспертиз используют дисперсионные модели [25]. Дисперсионные модели решают упрощенные формы уравнения адвекции-диффузии (1.2) и продуктивны для оценки распространения атмосферных примесей от точечных и линейных источников (промышленные предприятия, тепловые станции, дороги). По сравнению с гидродинамическими или химико-физическими моделями дисперсионные модели требуют небольшого объема исходных данных и имеют готовые стандартизированные реализации: классические «модель гауссова факела» (Gaussian plume) и «модель импульсного облака» или «пучка» (puff model) и более современные реализации, такие как AERMOD и CALPUFF [24–26].

В более сложных сценариях (нестандартная топография, сложные химические реакции, динамика водных объектов) дисперсионные модели не дают высокой точности. В этих случаях применяют гидродинамические или более сложные физические модели. Гидродинамические модели решают уравнения Навье-Стокса

вместе с уравнениями турбулентного переноса для атмосферы или водных объектов. Примеры таких моделей: WRF-Chem (Weather Research and Forecasting + chemistry), CMAQ (Community Multiscale Air Quality), HYSPLIT (Hybrid Single-Particle Lagrangian Integrated Trajectory), а для водной среды – MIKE 21 и Delft3D [27]. Эти модели способны давать детализированные поля распределения примесей, однако требуют значительных объемов исходных данных о метеоусловиях, инвентаризации эмиссий, характеристиках источников примесей и среды, а также высокой вычислительной мощности.

Преимущество физических моделей заключается в их физической обоснованности и возможности адаптации моделей к новым территориям или новым типам примесей без необходимости проведения дополнительных измерений содержаний примеси на территории. Вместе с тем применение физических моделей ограничено высокими требованиями к входным данным, чаще всего получаемых с помощью натурных измерений, и значительных вычислительных ресурсов. Сравнительные исследования показывают, что результаты моделирования поля примеси от одного источника могут существенно отличаться в зависимости от применяемой модели. Например, в работе [28] сравнивались дисперсионная модель CALPUFF и гидродинамическая модель CFD для токсичных газов в горной местности. Авторы отметили, что общие паттерны распределения примеси в целом согласуются между моделями, но точные результаты различаются, особенно вблизи источника, на порядок до сотен метров.

Таким образом, подходы на основе физического моделирования эффективны прежде всего для оценки воздействия крупных стационарных и протяженных источников примесей при наличии качественных данных о параметрах источников и среды. Однако в задачах, характеризующихся ограниченным числом наблюдений, разреженностью данных или отсутствием полной информации о характеристиках источников, применение физических моделей оказывается затруднительным. В этом случае обращаются к методам пространственной интерполяции, которые позволяют восстанавливать значения концентраций в недоступных точках пространства по имеющемуся набору наблюдений.

1.1.2 Методы пространственной интерполяции

Методы пространственной интерполяции решают задачу построения поля примеси $C(\boldsymbol{s})$ непосредственно по наблюдениям без явного учета физических механизмов формирования содержания примеси. Пространственная интерполяция представляет собой метод оценки значений некоторой величины (например, содержания примеси) в точках, где прямые измерения отсутствуют, на основе известных значений этой величины в ограниченном наборе точек наблюдений [29].

Пусть $C(\boldsymbol{s})$ – функция, определенная на пространственной области, и значения $C(\boldsymbol{s_i})$ известны в конечном числе точек наблюдений $\boldsymbol{s_i}$, $i = 1, \ldots, n$. Пространственная интерполяция представляет собой процедуру построения оценочной функции $\hat{C}(\boldsymbol{s})$, позволяющей аппроксимировать неизвестные значения $C(\boldsymbol{s})$ для любых точек $\boldsymbol{s}$ на основе известных пар $(\boldsymbol{s_i}, C(\boldsymbol{s_i}))$:

$$\hat{C}(\boldsymbol{s_0}) = \sum_{i=1}^{n} w_i(\boldsymbol{s_0}) C(\boldsymbol{s_i}), \quad \sum_{i=1}^{n} w_i(\boldsymbol{s_0}) = 1,$$

где $\hat{C}(\boldsymbol{s_0})$ – оценка значения величины в точке $\boldsymbol{s_0}$, $w_i(\boldsymbol{s_0})$ – весовой коэффициент, определяющий вклад наблюдения в точке $\boldsymbol{s_i}$ при оценке значения примеси в точке $\boldsymbol{s_0}$, зависящий от расстояния и/или ковариационной структуры между точками $\boldsymbol{s_i}$ и $\boldsymbol{s_0}$; $C(\boldsymbol{s_i})$ – наблюдаемое значение примеси.

Предполагается, что значение исследуемой величины изменяется непрерывно и существует зависимость между близко расположенными точками. Результатом интерполяции является поле распределения примеси, где каждой точке территории соответствует оцененное значение, полученное с помощью построенной аппроксимирующей функции.

В литературе традиционно выделяют две основные группы методов пространственной интерполяции: детерминистские (deterministic) и геостатистические (geostatistical) [30, 31].

Детерминистские методы не предполагают вероятностной модели и строят математическую поверхность по известным значениям в точках наблюдений, используя аналитические зависимости. К этой группе относятся:

- методы, основанные на пространственной близости точек (например, инверсное взвешивание расстояний – Inverse Distance Weighting, IDW), которые определяют значение в неизвестной точке как взвешенное среднее значений в известных точках, при этом вклад каждой известной точки уменьшается по мере удаления до нее;

- методы, основанные на сглаживании (сплайн-интерполяция, полиномиальные аппроксимации), которые формируют непрерывную и дифференцируемую поверхность, минимизирующую отклонения от исходных данных. В сплайн-интерполяции значение в неизвестной точке определяется как значение кусочно-гладкой функции, построенной таким образом, чтобы в известных точках она точно совпадала с измеренными значениями, а суммарная кривизна поверхности была минимальной. Полиномиальные аппроксимации – глобальные (global polynomial) или локальные (local polynomial) – подбирают аналитическую функцию заданной степени, описывающую общую тенденцию изменения содержания примеси на территории;

- методы на основе радиальных базисных функций (Radial Basis Functions, RBF) используют комбинацию радиальных ядер (например, гауссовых или многослойных), центрированных в точках наблюдений.

Детерминистские методы вычислительно просты, однако не дают оценки погрешности интерполяции и чувствительны к неоднородности сети наблюдений. Эта группа методов широко применяется преимущественно для первичного анализа пространственных данных и визуализации.

Геостатистические методы используются для более точного моделирования поверхностей, включая оценку ошибок и построение вероятностных параметров построенных поверхностей. В эту группу входит кригинг и его вариации (ordinary, universal, co-kriging и др.).

Геостатистические методы интерполяции основываются на стохастическом описании исследуемой величины $C(\boldsymbol{s})$, рассматриваемой как реализация случайного поля, обладающего пространственной корреляцией. В отличие от детерминистских методов, где веса $w_i(\boldsymbol{s_0})$ определяются геометрически

(например, через обратные взвешенные расстояния), в геостатистических методах эти веса рассчитываются с учетом ковариационной структуры данных [30, 31].

Веса $w_i(\boldsymbol{s_0})$ выбираются так, чтобы минимизировать дисперсию ошибки интерполяции при условии несмещенности оценки

$$\min_{w_i(\boldsymbol{s_0})} \mathrm{Var}\left[\hat{C}(\boldsymbol{s_0}) - C(\boldsymbol{s_0})\right], \qquad \sum_{i=1}^{n} w_i\,(\boldsymbol{s_0}) = 1.$$

Пространственная зависимость значений случайного поля описывается вариограммой:

$$\gamma(h) = \frac{1}{2}\mathrm{Var}\ [C(\boldsymbol{s}) - C(\boldsymbol{s}+h)],$$

где h – вектор расстояния между точками. Вариограмма характеризует степень изменчивости исследуемой величины в зависимости от расстояния и используется для расчета весов $w_i(\boldsymbol{s_0})$ в уравнениях кригинга.

На практике вариограмма аппроксимируется аналитической функцией, зависящей от расстояния. Например, для экспоненциальной модели

$$\gamma(h) = c_0 + c\,(1 - e^{-h/a}),$$

где параметры c_0, c, a определяются методом наименьших квадратов по экспериментальным данным.

Таким образом, кригинг представляет собой оптимальную линейную интерполяцию случайного поля, обеспечивающую минимальную дисперсию ошибки при известных статистических свойствах данных. Результатом геостатистической интерполяции является не только оценочное поле $\hat{C}(\boldsymbol{s})$, но и карта стандартной ошибки $\sigma(\boldsymbol{s})$, что позволяет исследователю оценивать точность получаемых результатов.

Эффективность методов пространственной интерполяции в значительной степени определяется пространственной неоднородностью распределения примесей, плотностью и репрезентативностью наблюдательной сети, а также качеством исходных измерений. Детерминистские методы чаще применяются на локальном уровне и преимущественно для визуализации данных, тогда как геостатистические методы позволяют учитывать пространственную корреляцию и

оценивать неопределенность результатов при региональных и мезомасштабных исследованиях [31].

Пример современного применения геостатистических методов пространственной интерполяции представлен в работе [32], где была выполнена реконструкция регионального распределения концентраций PM2.5 на территории Китая в период пандемии COVID-19. Другой пример представлен в работе [33], где для оценки приземных концентраций PM2.5 использовалась модификация кригинга, учитывающая направление ветра. В России кригинг был применен для оценки пространственного распределения загрязнения воздуха в г. Санкт-Петербург [34]. В этом исследовании использовался метод кокригинга, который использовал в качестве дополнительных переменных концентрации оксида углерода CO, оксида азота NO и диоксида азота NO_2 для более точной интерполяции концентрации взвешенных частиц PM10 в районах, где отсутствовали прямые измерения. Это позволило создать более детализированные карты загрязнения воздуха, несмотря на ограниченное количество данных о PM10.

Методы пространственной интерполяции имеют свои ограничения: они не позволяют локализовать источники загрязнения и чувствительны к неравномерности распределения точек наблюдений. В работах [32–34] для восстановления пространственного распределения PM2.5 методы интерполяции применялись с учетом вспомогательных переменных, что позволило получить непрерывные поля примеси и выявить пространственно-временные закономерности их изменения даже в условиях разреженных сетей наблюдений. При этом, несмотря на высокую детализацию полученных карт, подход не предусматривал явной идентификации вкладов отдельных источников примеси. Полученные результаты не позволяли напрямую восстановить параметры источников выбросов или механизмы формирования наблюдаемых концентраций.

Таким образом, применимость методов пространственной интерполяции в задачах моделирования полей примесей ограничена, особенно в условиях разреженных данных и сложной структуры источников. Это обусловливает необходимость привлечения подходов, которые позволяют учитывать нелинейные

зависимости и пространственные закономерности, недоступные классическим методам интерполяции.

1.2 Методы машинного обучения

В современных работах, посвященных пространственному моделированию распределения примесей в компонентах окружающей среды, активно применяются методы машинного обучения (machine learning, ML), позволяющие выявлять и моделировать сложные, как правило, нелинейные взаимосвязи между переменными-предикторами и содержанием примесей в компонентах окружающей среды:

$$\hat{C}(\boldsymbol{s}) = f(\boldsymbol{v}(\boldsymbol{s}), \boldsymbol{\theta}),$$

где $\boldsymbol{v}(\boldsymbol{s}) = (v_1(\boldsymbol{s}), v_2(\boldsymbol{s}), \ldots, v_p(\boldsymbol{s}))$ – вектор переменных-предикторов, $v_j(\boldsymbol{s})$, $j = 1, \ldots, p$ – значение j-го пространственного предиктора, характеризующее пространственные, антропогенные и природные характеристики территории, ассоциированные с точкой $\boldsymbol{s}$, а $f(\boldsymbol{v}(\boldsymbol{s}), \boldsymbol{\theta})$ – параметрическая функция, определяемая выбранной моделью машинного обучения, $\boldsymbol{\theta}$ – вектор ее параметров.

Первая группа методов – классические алгоритмы ML, такие как линейная регрессия, деревья решений и ансамблевые методы, которые представляют собой комбинации нескольких базовых ML моделей для повышения точности и устойчивости прогнозов. Среди ансамблей различают бэггинг, при котором модели обучаются на различных подмножествах исходного набора данных с последующим усреднением результатов; бустинг, при котором модели обучаются последовательно и таким образом корректируют ошибки предыдущих моделей; и стекинг, при котором прогнозы нескольких базовых моделей (и при необходимости исходные предикторы) используются для обучения метамодели, объединяющей результаты в итоговый прогноз. Вторая группа ML методов представлена нейросетевыми моделями – от классических многослойных перцептронов (MLP) до глубоких архитектур, включая сверточные (CNN), рекуррентные (RNN) и гибридные модели, объединяющие нейросетевые подходы с физико-ориентированными или статистическими моделями.

1.2.1 Классические алгоритмы машинного обучения

Классические алгоритмы машинного обучения используют предикторы, подготовленные заранее вручную и/или на основании экспертных знаний, и предполагают относительно простую архитектуру модели с малым числом параметров. В отличие от физических моделей и геостатистических методов интерполяции, классические алгоритмы ML не предполагают априорного физического или статистического описания распределения примеси, а строят зависимость исключительно на основе известных предикторов и наблюдаемых значений примеси. Такие методы в основном применимы к задачам, где объем данных ограничен.

Линейная регрессия (linear regression) – один из первых ML-методов, использованных в задачах пространственного моделирования полей примесей [35]. В современных работах линейная регрессия часто используется как базовая (baseline) модель для сравнения с более сложными подходами. Например, авторы [10] сравнили линейную пошаговую регрессию (stepwise linear regression) с различными методами регуляризации и ансамблевыми алгоритмами машинного обучения (Random Forest, Gradient Boosting/Bagging) для моделирования концентраций PM2.5 и NO_2 на территории Европы. Они показали, что применение линейной регрессии ограничено при наличии пространственной автокорреляции или сложных нелинейных взаимосвязей и что при большом объеме данных и разнообразии предикторов линейный метод уступает более сложным. В еще одной работе [36] множественная линейная регрессия использовалась для исследования зависимостей концентрации PM10 от метеорологических переменных (скорость ветра, направление, температура, влажность) в Стамбуле за зимний период. Авторы отметили, что линейная модель позволила выявить значимые связи (например, высокую концентрацию PM10 в зимний период), но подчеркнули ограниченность линейных моделей в условиях быстрого изменения метеоусловий.

Алгоритмы на основе деревьев решений (Decision Trees) и их ансамбли (Random Forest, Gradient Boosting) стали логичным развитием линейных моделей: они позволили учесть нелинейные взаимосвязи и взаимодействия признаков [37].

Деревья решений особенно продуктивны в задачах моделирования пространственного распределения примесей, так как они устойчивы к шумам и пропускам в данных [38], не требуют предварительной линеаризации признаков или строгого предположения о форме связи [39], хорошо справляются с набором предикторов различной природы (транспортные характеристики, землепользование, метеоусловия и др.) [40, 41]. Однако, как и в случае с линейной регрессией, игнорирование пространственной автокорреляции и неравномерности пространственного распределения наблюдений при построении моделей на основе деревьев решений может привести к смещенным результатам моделирования и завышенным оценкам качества моделей.

Такие модели как Support Vector Machine (SVM) и k-Nearest Neighbors (kNN) получили широкое распространение в задачах экологического моделирования в 2010-е гг. [42, 43]. Эти методы не требуют большого набора данных, но позволяют работать с нелинейными зависимостями и многомерным набором предикторов за счет преобразования пространства признаков (в случае SVM) или применения локального правила ближайших соседей (в случае kNN). Например, в исследовании по оценке концентраций PM2.5 в провинции Хубэй, Китай применялась модифицированная SVM-модель, где авторы указывают, что «SVM … показывает преимущества при малых выборках, при высокоразмерных данных» [44]. В отличие от деревьев решений, методы этого класса не дают прямой интерпретации влияния предикторов, но нередко обеспечивают более точную аппроксимацию при ограниченном числе переменных.

Таким образом, классические алгоритмы ML хорошо подходят для задач с ограниченным числом наблюдений и заранее подготовленными предикторами, однако при сложных пространственных зависимостях и больших объемах данных их точность и способность моделировать нелинейные связи снижается. В таких ситуациях применяются нейросетевые и гибридные архитектуры, способные учитывать как пространственные, так и физические закономерности формирования полей примесей.

1.2.2 Нейросетевые модели и гибридные архитектуры

Искусственные нейронные сети (ИНС) представляют собой класс нелинейных моделей машинного обучения, которые способны аппроксимировать сложные зависимости между входными и выходными переменными [45]. В отличие от линейных моделей, нейросети могут выявлять нелинейные закономерности и взаимодействия факторов, что делает их перспективным инструментом для задач построения пространственного распределения примесей [46].

Архитектура нейросети определяется количеством слоев, нейронов и функциями активации, а также типом связей между слоями (полносвязные, сверточные, рекуррентные и др.), а процесс обучения заключается в настройке весов модели с использованием итеративных оптимизационных процедур, например алгоритма обратного распространения ошибки [47]. Обученная ИНС отражает закономерности, присутствующие в данных, и может рассматриваться как функциональный аналог традиционной модели [48]. Однако, в отличие от традиционных моделей, зависимости между переменными в ИНС не задаются явно. Нейронные сети представляют собой типичный пример эмпирического подхода к исследованию, т. е. «черного ящика», хотя в последние годы активно развиваются методы объяснимого искусственного интеллекта (XAI), позволяющие частично интерпретировать результаты ИНС [49].

Одной из наиболее популярных ИНС, используемых в экологических исследованиях, является многослойный персептрон (MLP). Благодаря своему широкому использованию этот тип сети хорошо развит и показал свою высокую продуктивность: обычно MLP превосходит линейную регрессию, деревья решений и классические методы интерполяции [50, 51]. Структура сети MLP описывается несколькими значениями, связанными с количеством нейронов в слоях: входной слой – скрытый слой (или несколько скрытых слоев) – выходной слой. Персептроны широко используются в исследованиях распределения химических элементов в почве [50–55]. Многие исследователи использовали MLP для

цифрового картирования почв [54, 55] и доказали превосходство моделей MLP над геостатистикой.

Современные ИНС включают глубокие сети (deep learning), сверточные (convolutional neural networks, CNN) и рекуррентные (recurrent neural networks, RNN) архитектуры. Сверточная нейронная сеть, предложенная Яном Лекуном в 1988 г., изначально задумана для обработки изображений и пространственно структурированных данных [56]. CNN способны автоматически извлекать пространственные признаки из входных данных за счет применения сверточных фильтров, что делает их особенно подходящими для анализа географических и экологических полей (например, карт концентраций загрязняющих веществ, цифровых моделей рельефа, спутниковых снимков и т. п.). Благодаря этому свойству, сверточные архитектуры активно используются для реконструкции и прогнозирования пространственного распределения загрязнений воздуха и воды, оценки качества среды, а также для обработки данных дистанционного зондирования. CNN могут превосходить традиционные геостатистические и MLP-модели за счет более глубокого и адаптивного представления пространственных закономерностей (например, [57]).

Рекуррентные нейронные сети и их модификации, в частности, LSTM (Long Short-Term Memory) и GRU (Gated Recurrent Unit), предназначены для анализа временных зависимостей. В экологическом контексте они используются для моделирования динамики загрязнений, прогнозирования временных рядов концентраций и оценки влияния метеорологических факторов. Поскольку данные наблюдений за примесями часто представляют собой пространственно-временные последовательности, RNN и их вариации позволяют учитывать не только пространственные, но и временные закономерности, обеспечивая более точное моделирование процессов переноса и рассеивания примесей [58].

Современные исследования активно развивают гибридные архитектуры, которые объединяют преимущества различных типов моделей. Например, CNN-LSTM-модели сочетают способность CNN к извлечению пространственных признаков с возможностью LSTM моделировать временные зависимости, что

делает их особенно эффективными для задач моделирования пространственно-временного распределения примесей. Также развиваются подходы, комбинирующие нейронные сети с физико-химическими моделями или геостатистическими методами (например, CNN-Kriging, Physics-informed Neural Networks), что позволяет повысить интерпретируемость и физическую согласованность получаемых результатов ([5]).

Количество данных, необходимое для обучения нейронной сети, зависит от сложности модели, размерности признакового пространства и степени вариативности данных. При недостатке данных сеть склонна к переобучению и фактически приближается по поведению к линейной модели. Архитектуры без скрытых слоев и без нелинейных активаций эквивалентны линейной регрессии и не способны аппроксимировать нелинейные зависимости между переменными. Поэтому в задачах с ограниченным объемом наблюдений обычно используют классические ML модели, упрощенные архитектуры, регуляризацию или комбинированные подходы с включением априорной информации.

1.3 Подход Land Use

Моделирование пространственного распределения примеси предполагает подбор предикторов, которые максимизируют прогностическую способность модели. Эта идея реализуется в подходе Land Use (LU), который был разработан и впервые применен в 1990-х гг. в Европе для оценки персональных экспозиций населения к автотранспортным выбросам в рамках крупных эпидемиологических исследований [8]. Первоначально метод назывался Regression mapping (регрессионная картография), в последующих работах закрепился современный термин Land Use Regression (LUR). Идея LU подхода заключается в построении модели, описывающей пространственные взаимосвязи между измеренными содержаниями примесей и пространственными характеристиками территории.

LU модели позволяют прогнозировать содержание примеси в точках, где непосредственные измерения не проводились, что делает подход особенно ценным при малом числе наблюдений. При этом в отличие от геостатистических методов

интерполяции (например, кригинга), LU подход учитывает пространственное взаиморасположение источников примесей по отношению к точкам наблюдений [8–18, 59, 60]. Это обеспечивает более достоверное описание пространственной неоднородности загрязнения, характерной в том числе для городских территорий, где существенную роль играют локальные источники примесей (автотранспорт, строительные площадки, карьеры и т. п.).

1.3.1 Термин Land Use в отечественной и зарубежной практике

Термин land use («землепользование») широко используется в экологических и градостроительных исследованиях. В отечественной нормативной системе этот термин закреплен в законодательных документах, определяющих правовые аспекты использования земель. В Градостроительном кодексе Российской Федерации землепользование рассматривается в контексте градостроительного зонирования, определяющего виды разрешенного использования земельных участков в пределах границ соответствующей территориальной зоны [61]. Федеральный закон «О введении в действие Земельного кодекса Российской Федерации» (от 08.08.2024 г. № 307-ФЗ) подчеркивает связь использования земель с видами разрешенного использования и предписывает «обязанность использовать земельные участки по целевому назначению» [62]. ГОСТ Р 59055-2020 «Охрана окружающей среды. Земли. Термины и определения» устанавливает «термины и определения… при использовании земель», но сам термин «землепользование» в тексте ГОСТ не определяется [63].

В зарубежной практике термин land use также имеет нормативное закрепление, но преимущественно как функционально-экологическая характеристика территории, используемая в том числе в системах классификации земных покровов и картографии. Так, U.S. Environmental Protection Agency (EPA) определяет land use как форму экономической и культурной деятельности (например, сельскохозяйственной, жилищной, промышленной, горнодобывающей, рекреационной), осуществляемую на конкретной территории [64]. В классификации U.S. Geological Survey (USGS) land use разграничивается от land

cover («земельный покров» – физическое состояние поверхности Земли, например, лес, вода, пашня, застроенная территория) и описывается как способ адаптации поверхности Земли под нужды общества, включающий такие категории использования, как урбанизированные территории (urban), сельскохозяйственные угодья (agricultural), леса (forest), водно-болотные угодья (wetland) и др. [65]. Аналогичные определения приводятся в глоссариях по градостроительному планированию, например, в Barre City Land Use Glossary (2023), где под land use понимается способ адаптации или модификации физического мира для человеческих нужд [66].

1.3.2 Схема Land Use подхода

Построение модели LU включает несколько этапов.

1. Измерение содержания примесей в точках наблюдений на исследуемой территории.

На первом этапе осуществляют выбор территории с источниками примесей и координатами n точек наблюдений, в которых измерены содержания примесей. Таким образом, для каждой примеси каждая i-я точка наблюдения характеризуется координатами $\boldsymbol{s_i} = (x_i, y_i)$ и измеренными содержаниями примесей $C_i = C(\boldsymbol{s_i})$, $i = 1, \ldots, n$ – порядковый номер точки наблюдения.

Существует несколько стратегий выбора точек наблюдений на исследуемой территории. Категориальный (стратифицированный) подход предполагает предварительное разбиение территории на категории с предположительно разным уровнем загрязнения, например, «городской фон», «город — интенсивное движение» и «сельский район». Критериями для отнесения точки наблюдения к категории служат интенсивность автомобильного трафика (например, для «городского фона» – не более 3000 автомобилей/сутки в радиусе 50 м) и тип застройки («открытая местность» или «уличный каньон»). После классификации исследователи определяют соотношение точек наблюдений между этими категориями [8, 67]. Формализованный (целевой) подход использует алгоритмы выбора точек наблюдений, которые учитывают пространственные переменные,

закладываемые в будущую модель. Например, в Канаде [68] для размещения точек использовались данные о транспортной сети и местах проживания когорты исследования. В работе устанавливали больше пробоотборников в районах с предполагаемой высокой пространственной неоднородностью содержания примеси (например, по плотности дорог) и в зонах с наибольшей плотностью населения.

Общий для обоих подходов принцип заключается в том, чтобы подобрать точки наблюдений таким образом, чтобы охватить весь диапазон возможных содержаний примесей в пределах исследуемой территории. От репрезентативности данных наблюдений зависит способность модели прогнозировать содержания примесей в тех точках, где измерения не проводились.

Выбор числа точек наблюдений ограничен физическими и материальными возможностями исследователя, наличием необходимого измерительного оборудования. Универсальной методологии для определения оптимального числа точек наблюдений не существует, и их количество в различных исследованиях варьируется от 20 до 150 [12, 68–70]. Малое число точек наблюдений (менее 40) может привести к неточным оценкам содержания примесей. В то же время увеличение их числа также не всегда приводит к статистически значимому улучшению модели. Оптимальное количество точек должно определяться местными особенностями, ожидаемой вариацией содержания примесей и размерами исследуемой территории. Для моделирования пространственного распределения примесей в крупном городе считается целесообразным использовать от 40 до 80 точек отбора проб [71]. Такое количество точек определяется необходимым объемом данных для статистического анализа и не зависит от размера города.

Для построения модели используют данные о содержании примесей, полученные в ходе мониторинга или, более предпочтительно, скрининга.

2. Сбор данных о пространственных особенностях территории и построение пространственных переменных

На следующем этапе построения LU-модели необходимо собрать доступные пространственные данные, оценить их возможную взаимосвязь с содержанием примесей и построить набор переменных-потенциальных предикторов, отражающих эти взаимосвязи. Каждая переменная при этом представляет гипотезу о направлении и форме связи географического фактора с содержанием примеси.

Для исследуемой территории создается геоинформационная модель – географическая база данных, включающая информацию о транспортной сети (типах дорог, интенсивности движения), землепользовании (жилая, коммерческая, промышленная, сельскохозяйственная застройка, парки, водные объекты), топографии (высота над уровнем моря, рельеф), плотности населения и застройки, расположении промышленных предприятий и других потенциальных источников примесей. Источниками данных могут служить официальные картографические материалы, открытые данные (например, спутниковые снимки Google Earth и специализированные геоинформационные базы данных).

Далее для каждой точки наблюдения рассчитывается ряд пространственных переменных. В LU подходе основным приемом для вычисления пространственных переменных является построение буферов (buffers) – круговых окрестностей точек наблюдений заданного радиуса, внутри которых производится количественная оценка различных факторов землепользования: длин дорог, площадей парковых, жилых и промышленных зон. Некоторые примеры пространственных переменных:

- на основе данных о дорожно-транспортной сети, например, расстояние от точки наблюдения до ближайшей дороги; плотность дорог всех типов внутри круговых окрестностей точки наблюдения с различными радиусами от 100 до 1000 м; общая длина дорог внутри круговых окрестностей точки наблюдения радиусами от 100 до 1000 м; общая длина главных дорог города внутри круговых окрестностей точки наблюдения с радиусами от 100 до 1000 м.

- на основе данных об интенсивности трафика, например, интенсивность движения транспорта или отдельных видов транспорта внутри круговых окрестностей точки наблюдения.

- на основе географических и топографических данных, например координаты точки наблюдения; высота над уровнем моря; расстояние до моря или других крупных географических объектов;

- на основе данных о типах землепользования, например площади, занятые под жилую, коммерческую, промышленную застройку, а также под парки и водные объекты в круговых окрестностях различного радиуса;

- на основе демографических данных, например, плотность населения в круговых окрестностях точки наблюдения различного радиуса.

Например, в работе [8] LU модели, которые были построены с четырьмя различными типами пространственных переменных, включая тип дороги, интенсивность движения, землепользование и физическую географию, относительно хорошо смогли оценить содержание NO_2.

Радиусы круговых окрестностей точек наблюдений и состав переменных подбираются эмпирически, исходя из максимизации объясненной дисперсии содержания примесей. Выбор радиусов является критичным для качества и пространственного разрешения модели и должен быть основан на информации о вероятном рассеянии примеси и предположениях о влиянии на нее изучаемых предикторов. Радиусы могут быть от самых малых (например, нескольких метров – насколько позволяет пространственное разрешение данных) до очень больших (но обычно не превышают 1–2 км, поскольку с увеличением радиуса снижается интерпретируемость предикторов, в том числе теряется предметный смысл переменных с большим радиусом).

Поскольку априори неизвестно, какие переменные имеют наибольшую связь с содержанием примеси, и неизвестно, какой радиус круговой окрестности следует выбрать для наилучшего описания этой связи, изначально рассчитывается большой набор пространственных переменных – потенциальных предикторов модели. Обычно исследователи способны преобразовать собранную информацию в 50–150 различных пространственных переменных. Например, исследование [72] включало 55 потенциальных предикторов, исследование [73] – 140 потенциальных предикторов.

3. Отбор предикторов

Полученный набор потенциальных пространственных переменных подвергается анализу для выявления небольшой группы предикторов, наиболее тесно связанных с измеренным содержанием примеси. Основное требование к создаваемым моделям пространственного распределения примесей заключается в достижении максимального качества аппроксимации (по значениям коэффициента детерминации R^2, среднеквадратической ошибки и др.) при минимальном числе переменных-предикторов. С этой целью применяется статистический отбор предикторов, направленный на исключение пространственных переменных с низкой корреляцией с целевой переменной, пространственных переменных, сильно коррелирующих между собой (мультиколлинеарных), статистически незначимых пространственных переменных, пространственных переменных, противоречащих физическому смыслу модели и т. д.

В практике LU-моделирования для отбора предикторов применяется несколько подходов:

- пространственные переменные ранжируются по величине парной корреляции с измеренными содержаниями примесей [74],

- применяются пошаговые процедуры регрессии, например, Forward stepwise regression – переменные добавляются к модели последовательно, если прирост R^2 значим по F-критерию, Backward elimination – переменные исключаются, если их *p-value* превышает 0,1 [9].

- выполняется полный перебор всех возможных комбинаций пространственных переменных (при малом числе кандидатов) с выбором модели с наилучшим значением *Adjusted* R^2, *AIC* или *BIC* [9].

- применяются регуляризационные методы, такие как LASSO, Ridge и Elastic Net, позволяющие автоматически исключать неинформативные переменные при обучении модели [9, 10].

- для оценки мультиколлинеарности – наличия сильной линейной корреляции между переменными регрессионной модели – используется Variance Inflation Factor (VIF) [75].

В итоговую модель обычно включают от 3 до 8 наиболее значимых предикторов. Увеличение числа предикторов может повысить качество модели, однако затрудняет ее предметную интерпретацию.

4. Построение модели

На следующем шаге строится модель, описывающая взаимосвязь между измеренными содержаниями примесей и полученным набором предикторов. Классический LU подход предполагает использование множественной линейной регрессии в качестве интерполятора:

$$C_i = C(\boldsymbol{s_i}) = \beta_0 + \sum_{j=1}^{p} \beta_j v_j(\boldsymbol{s_i}) + \varepsilon_i, \quad (1.3)$$

где $i = 1, \ldots, n$, n – число наблюдений,

$j = 1, \ldots, p$ – число предикторов,

$C_i = C(\boldsymbol{s_i})$ – зависимая переменная, например, содержание примеси в точке наблюдения $\boldsymbol{s_i}$,

v_1, v_2, ..., v_p – независимые переменные (предикторы), например, v_1 – суммарная длина автомобильных дорог в окрестности точки наблюдения с радиусом 100 м, v_2 – площадь промышленных территорий в окрестности точки наблюдения с радиусом 500 м, v_3 – доля зеленых насаждений в окрестности точки наблюдения с радиусом 1000 м, и т. д.,

β_0 – свободный член, отражающий базовое (фоновое) содержание примеси при нулевых значениях v_1, v_2, ..., v_p,

β_1, β_2, …, β_p – коэффициенты регрессии, определяющие направление и силу влияния соответствующего предиктора на содержание примеси,

ε_i – ошибка, учитывающая случайную вариацию, которая не объясняется включенными в модель предикторами.

В матричной форме модель запишется как:

$$\boldsymbol{C} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1.4)$$

где $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ – матрица предикторов, $\boldsymbol{C} = (C(\boldsymbol{s_1}), \ldots, C(\boldsymbol{s_n}))^{\top}$ – вектор наблюдаемых значений примеси, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\top}$ – вектор коэффициентов.

Оценка коэффициентов осуществляется, как правило, методом наименьших квадратов:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\boldsymbol{C} \qquad (1.5)$$

Построенная модель определяет оценку содержания примеси в произвольной точке территории:

$$\hat{C}(\boldsymbol{s}) = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j \, v_j(\boldsymbol{s}).$$

5. Валидация модели

Заключительный этап построения модели Land Use – оценка качества и обобщающей способности модели. В простейшем варианте модель проверяется с помощью перекрестной проверки (cross-validation), при которой исходные данные многократно разделяются на обучающее и тестовое подмножества. В качестве базовых схем могут применяться разделение на две части (hold-out), leave-one-out cross-validation (LOOCV) или k-fold cross-validation, в зависимости от объема данных и пространственной корреляции между наблюдениями (см. подробнее в разделе 1.4.2). Цель валидации – получение несмещенной оценки точности прогноза и проверка устойчивости модели к вариациям данных.

1.3.3 Применение Land Use моделей для построения пространственного распределения примесей

Полученные модели применяются для построения пространственного распределения примесей в различных компонентах окружающей среды, особенно в условиях сложной городской структуры, где существенную роль играют локальные источники примесей: автомобильные дороги, промышленные зоны, строительные площадки и др. [8, 10–18, 59, 60]. В большинстве исследований LU подход применялся для оценки содержания атмосферных примесей: оксидов азота (NO_2, NO) [71, 74, 76], взвешенных частиц (PM2.5, PM10) [17, 74], диоксида серы [77], нефтепродуктов (BTEX: бензол, толуол, этилбензол, ксилолы) [76], летучих органических соединений (VOC) [78], сажи [79] и др. Существенно реже встречаются исследования, посвященные оценке распределения примесей в

почвенном или снеговом покрове [80, 81]. Между тем именно эти компоненты среды чувствительны к локальным особенностям землепользования и могут быть эффективно описаны в рамках LU подхода. В России результаты LU моделирования представлены, например, работами [59, 60, 80].

В современных исследованиях подход LU продуктивно дополняется ML моделями для повышения точности прогноза [14, 15]. Помимо классических линейных моделей регрессии, активно развиваются гибридные и ML версии LU, сочетающие идеологию LU с современными алгоритмами анализа данных. Такие подходы позволяют учитывать нелинейные и сложные пространственные зависимости между источниками и содержанием примесей. В последние годы предложен ряд гибридных моделей, объединяющих LUR с кригинговыми методами, случайным лесом (Random Forest), бустингом (XGBoost), глубокими нейронными сетями, многослойными перцептронами и сверточными нейронными сетями [14, 15, 17]. Например, работы [17, 82, 83] показывают, что использование алгоритмов машинного обучения значительно повышает точность прогнозов по сравнению с классическим LUR. Похожие результаты получены в исследованиях [14–17, 85], где продемонстрировано, что нейросетевые и ансамблевые версии LUR обеспечивают более высокое значение коэффициента детерминации R^2 и лучшую пространственную согласованность результатов.

LU подход позволяет получать оценки распределения примесей для значительной по площади территории (целый город) при относительно небольших материальных и временных затратах [13, 86–88]. Еще одной перспективной для решения с помощью LU подхода задачей может быть оценка вклада различных по природе источников в пространственное распределение примесей.

1.4 Подходы к оценке качества и устойчивости моделей

1.4.1 Основные понятия в задаче оценки моделей

При построении моделей прогноза и/или реконструкции чаще всего встречается разбиение используемого набора данных на следующие три

подмножества: обучающее подмножество (training set), валидационное подмножество (validation set), тестовое подмножество (test set) [89].

Обучающее подмножество (training set) – это подмножество набора данных, которое используется для обучения модели. Обучение – процесс подбора параметров (весов и смещений), составляющих модель (рисунок 1.1). Во время обучения модель итеративно обрабатывает примеры из обучающего подмножества и корректирует параметры. При этом каждый пример из обучающего подмножества используется от нескольких раз до миллиардов раз [90].

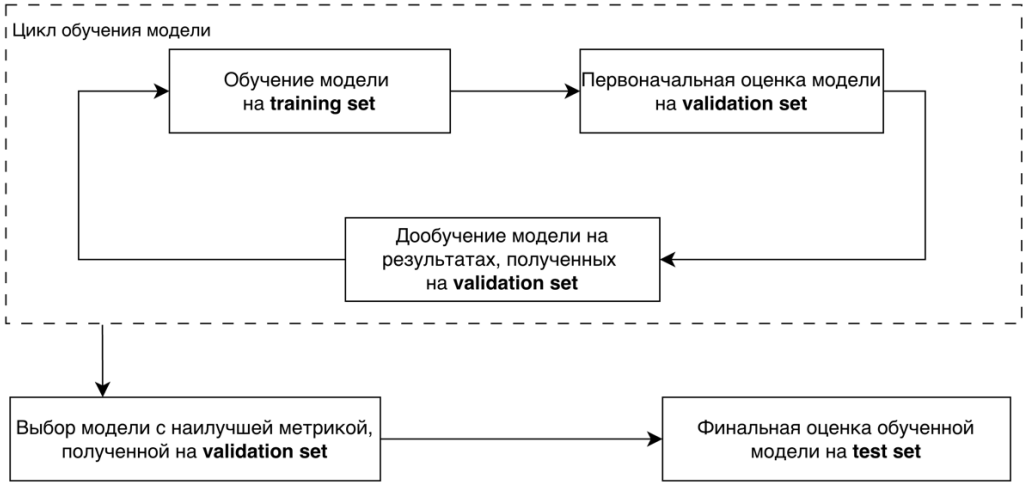


Рисунок 1.1 Обучение, валидация и тестирование модели

Валидационное подмножество (validation set) – это подмножество набора данных, которое используется для первоначальной оценки обученной модели. Обычно модель оценивают на валидационном подмножестве несколько раз до проведения финального тестирования на тестовом подмножестве [90]. Валидационное подмножество служит для настройки гиперпараметров (т. е. архитектуры) модели. Иногда его называют *development set*/*dev set* или *tuning set*, что в дословном переводе на русский язык означает «*отладочное подмножество*», но чаще переводится как «*валидационное подмножество*», чтобы соблюсти терминологическую согласованность [91].

Тестовое подмножество (test set) — это подмножество набора данных, зарезервированное для финального тестирования обученной модели. Поскольку тестовое подмножество лишь косвенно связано с процессом обучения, ошибка,

вычисленная на тестовом подмножестве (test loss), является менее смещенной и более честной метрикой по сравнению с ошибкой, вычисленной на обучающем или валидационном подмножестве [90].

На практике часто возникает путаница между терминами «validation set» (или «development set», «tuning set») и «test set» (таблица 1.1). Уже в первых статьях по глубокому обучению, опубликованных в престижных медицинских журналах, наблюдается несогласованность в терминологии, связанной с разбиением наборов данных на подмножества [92]. Ярким примером служит статья [93]. В ней термины «tuning set» и «validation set» используются для обозначения подмножеств, которые в работах по ML обычно известны как «validation set» и «test set» соответственно. При этом у тех же авторов в другой работе встречается пример использования общепринятой терминологии [94].

Таблица 1.1 Примеры использования терминологии в статьях по ML

Терминология			Ссылки
Обучение модели, подбор параметров	Дообучение, настройка гиперпараметров модели	Оценка модели	
train/training	validation	test	[94, 96]
train/training	tuning	test	[97]
train/training	development	test	[98, 99]
train/training	tuning	validation	[93, 100]
train/training	development	validation	[92]

Аналогично, исследование [95] применяет термины «development set» и «validation set» вместо привычных «training/validation set» и «test set», как это делается в немедицинских публикациях [92]. Авторам обзорных исследований приходится сталкиваться с такой несогласованностью напрямую, интерпретируя различные термины и выбирая, какую номенклатуру использовать в своей работе. Эта вариативность терминологии может легко вводить исследователей в заблуждение. Наблюдается постепенный сдвиг в сторону стандартизации: например, в работе [96] используется традиционная номенклатура: «training set»,

«validation set», «test set», при этом каждый термин четко определен в приложении, что облегчает воспроизводимость и интерпретацию результатов.

1.4.2 Подходы к валидации моделей

Оценка качества модели – необходимое условие для вывода о ее обобщающей способности. В задачах пространственного моделирования классические схемы валидации, такие как hold-out и random k-fold, часто дают оптимистичную оценку из-за пространственной автокорреляции наблюдений. Для получения честной оценки предсказательной способности модели следует использовать стратегии, учитывающие пространственную структуру данных [101, 102].

Подходы к валидации моделей делятся на две крупные группы: resampling-free – подходы, при которых выполняется разовое разбиение набора данных (hold-out validation), и ресэмплинговые (resampling-based) подходы, которые предполагают многократное разбиение набора данных (cross-validation) (рисунок 1.2).

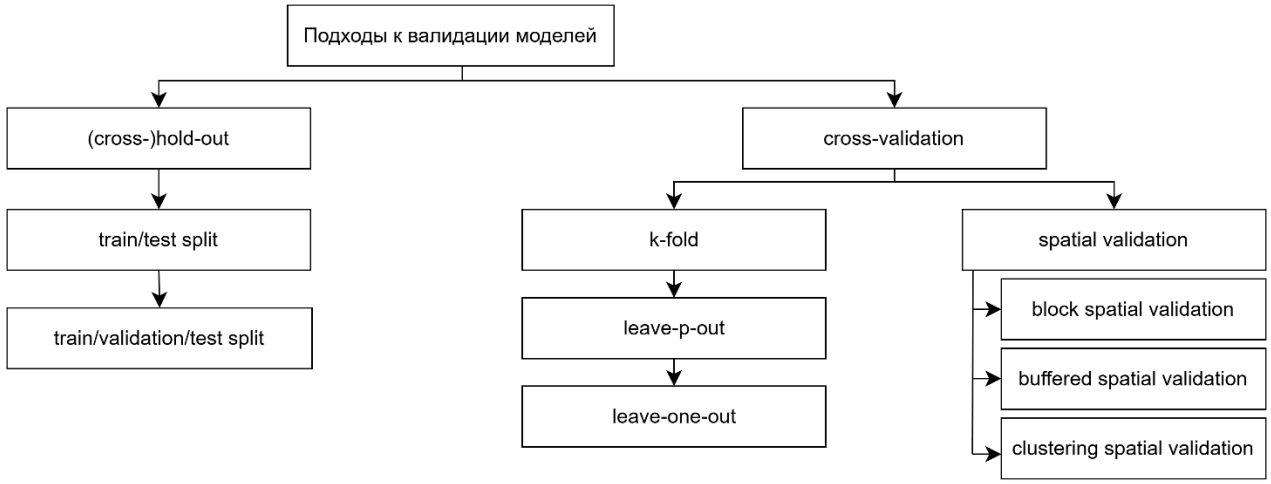


Рисунок 1.2 Подходы к валидации моделей

Метод *hold-out* (train/test split) – простая схема валидации, при которой набор данных разбивается на тренировочное и тестовое подмножества (например, в соотношении 70/30). Преимуществом этого подхода является простота и скорость, однако при малых выборках высока вариативность оценки, а также существует риск смещения, если тестовое подмножество нерепрезентативно по пространству

или значениям предикторов. Hold-out уместен для быстрых предварительных проверок на достаточно больших и пространственно равномерных наборах данных.

Более надежную проверку, особенно при использовании сложных моделей, требующих подбора параметров (например, регуляризованных регрессий или нейросетей), обеспечивает вариация классического hold-out (train/validation/test split), при которой выборка делится на три подмножества: обучающее, валидационное, тестовое. Однако при малых выборках уменьшение объема данных в обучающем подмножестве может ухудшить устойчивость модели.

Когда набор данных небольшой, и выделение отдельного валидационного подмножества нецелесообразно, используют *k-fold cross-validation* (CV). При *k*-fold cross-validation набор данных делится на k частей, обычно k = 5–10; модель обучается на k-1 частях, а оставшаяся часть служит для валидации; процедура повторяется k раз, и вычисляются усредненные метрики качества. Частный случай *k*-fold cross-validation – *Leave-One-Out* (LOO) cross-validation, при котором $k = n$, т. е. на каждой итерации из обучающего подмножества исключается одно наблюдение и используется для тестирования модели. LOO эффективен при очень малых выборках, но часто дает оптимистичные результаты при наличии пространственной автокорреляции. В литературе по LUR LOO традиционно использовался как стандарт оценки, однако исследования указывают на необходимость осторожной интерпретации результатов в пространственно зависимых данных [102, 103].

Еще один вариант *hold-out* валидации – cross-hold-out или repeated hold-out, представляющий собой многократное повторение (например, 100 или 1000 раз) стандартной процедуры *hold-out* [104]. При классическом *hold-out* модель обучается и оценивается один раз, при этом оценка качества модели может сильно зависеть от случайного разбиения. При cross-hold-out для повторения процедуры каждый раз создается новое случайное разбиение на обучающее и тестовое подмножества, результатом оценки модели является не одно значение метрики, а их распределение. Это позволяет получить более надежную оценку качества модели, усредненную по всем разбиениям.

Классические схемы валидации не учитывают пространственную зависимость точек наблюдений, что приводит к «утечке информации» между обучающим и тестовым подмножествами. Для пространственных данных применяют пространственную кросс-валидацию (*spatial CV*) – семейство модификаций *k*-fold, где разбиение формируется с учетом координат точек наблюдений. Основные варианты spatial CV включают:

- spatial *k*-fold (blocking / tiling) валидация, при которой территория разбивается на *k* пространственных блоков (регулярной сеткой или кластеризацией точек). На каждой итерации *k*-й блок исключается из обучения и используется в качестве тестового подмножества. Такой подход снижает влияние автокорреляции, поскольку точки из соседних областей не попадают в разные подмножества. blocking CV обеспечивает более реалистичную оценку способности модели обобщать на новые участки пространства [103].

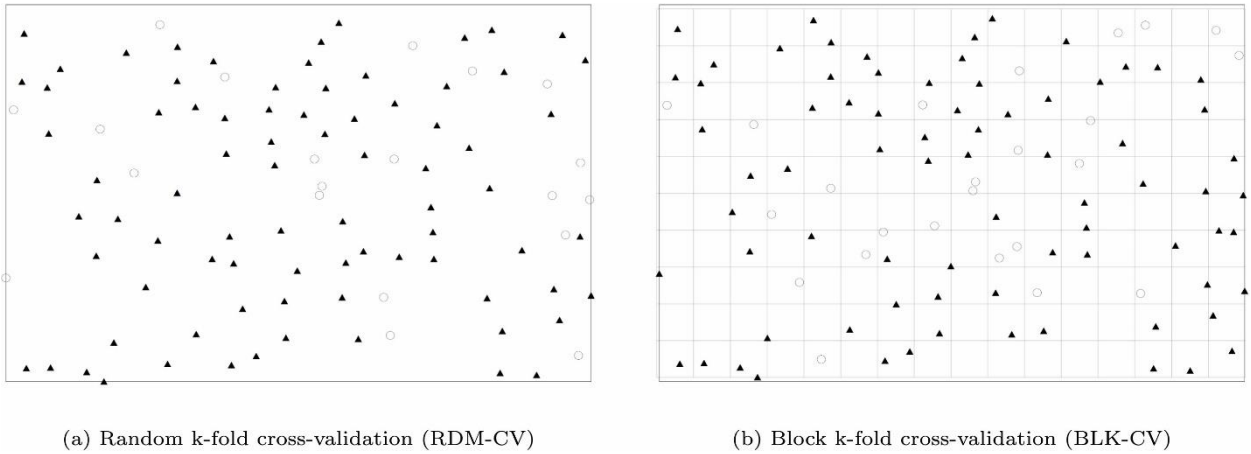
- buffered CV / buffered-LOO (distance-based exclusion) валидация, при которой для каждой тестовой точки исключаются из обучения все точки, находящиеся на расстоянии менее *r*:

$$Train = \{\boldsymbol{s_j}: d(\boldsymbol{s_i}, \boldsymbol{s_j}) > r\},$$

где $d(\boldsymbol{s_i}, \boldsymbol{s_j})$ – евклидово расстояние между точками, r – радиус исключения, определяемый по вариограмме или экспертно.

Такой подход устраняет влияние ближайших соседей и обеспечивает независимость обучающего и тестового подмножеств, особенно при высокой локальной автокорреляции [102].

- *clustering* / *leave-location-out* валидация (еще встречается название *grouped CV* или *spatial group CV*), которая используется в случае, если данные содержат повторные измерения на одних и тех же станциях или локациях. В тестовое подмножество включаются все точки, принадлежащие одному кластеру, оставшиеся кластеры используются для обучения (рисунок 1.3). Это позволяет проверить способность модели предсказывать значения в новых локациях, а не только для новых наблюдений [101].



(a) Random k-fold cross-validation (RDM-CV) (b) Block k-fold cross-validation (BLK-CV)

Рисунок 1.3 Случайная *k*-fold кросс-валидация и *k*-fold кросс-валидация по блокам [105]

Сравнительные исследования показывают, что при существенной пространственной автокорреляции блоковая или буферная CV дают более консервативную и реалистичную оценку предсказательной способности по сравнению с классическим random *k*-fold [101, 102].

1.4.3 Метрики качества моделей

Для оценки качества модели традиционно используются метрики ошибок (например, средняя абсолютная ошибка, среднеквадратическая ошибка) и метрики точности (коэффициент корреляции и т. д.). Различные типы метрик точности (или ошибок) имеют смысл расстояния (или обратного расстояния) между наборами наблюдаемых и предсказанных данных. Метрики качества можно разделить на два класса. Первый класс таких метрик, которые можно назвать одиночными (соло метриками), оценивает лишь один аспект соответствия или расхождения, специфичный для данной метрики [106–108]. Примерами таких метрик являются среднеквадратическая ошибка и коэффициент корреляции. Второй класс метрик качества представлен интегральными метриками, которые объединяют оценки из нескольких одиночных метрик. Однако интегральные метрики не позволяют определить, какой именно аспект соответствия или расхождения в большей степени повлиял на полученное значение. Чтобы определить этот аспект, необходимо вернуться к одиночным метрикам.

Для удобства восприятия полученные метрики можно визуализировать. Один из успешных примеров совместного использования скалярных метрик – построение диаграммы Тейлора [109, 110], которая широко используется в оценке и тестировании моделей [106, 111, 112]. Диаграмма Тейлора объединяет в себе три показателя: стандартное отклонение, среднеквадратическую ошибку и коэффициент корреляции. Стандартное отклонение позволяет оценить вариативность предсказанных данных, которая, как предполагается, не должна превышать вариативность наблюдаемых данных. По коэффициенту корреляции возможно оценить согласованность изменчивости между наблюдаемыми и предсказанными наборами данных. Среднеквадратическая ошибка показывает среднеквадратическое расхождение между наблюденными и предсказанными значениями.

Абсолютные значения упомянутых метрик пригодны для сравнения конкурирующих моделей, но в ситуации, когда рассматривается единственная модель, такие метрики оказываются малоинформативными. Использование комбинации различных метрик, безусловно, дает дополнительную информацию о качестве модели, но, поскольку эти метрики коррелированы друг с другом, их совместный анализ не всегда приводит к существенному расширению представлений о предсказательной способности модели. Более того, сами значения метрик, полученные на конкретном наборе данных, не дают ответа на вопрос о том, насколько они устойчивы к вариациям в данных и значимы в вероятностном смысле.

1.4.4 Методы ресэмплинга

В рамках статистического вывода в соответствии с идеей тестирования гипотез задача оценки качества модели сводится к вычислению *p-value* как меры уверенности/неуверенности (в зависимости от используемой статистики) в достаточности качества полученной модели. Применение традиционно используемых методов статистического вывода, основанных на асимптотических приближениях, связано с рядом ограничений. Асимптотические тесты опираются

на предположения о виде распределений данных или статистик, которые на практике почти всегда нарушаются [113, 114]. В результате использование асимптотических тестов может приводить к смещенным оценкам *p-value* и, как следствие, некорректным выводам о качестве моделей, особенно в условиях малого числа наблюдений.

Альтернативой асимптотическим подходам являются непараметрические методы, в частности методы ресэмплинга (resampling methods), которые не требуют строгих предположений о распределении данных. К методам ресэмплинга относятся пермутационные тесты (permutation tests), бутстрэп (bootstrap) и jackknife. Их объединяет общая вычислительная идея – формирование распределения интересующей статистики путем повторных случайных преобразований исходного набора данных [115, 116]. Эти методы позволяют оценивать неопределенность метрик качества, строить доверительные интервалы и вычислять *p-value* на основе эмпирического распределения статистики [113, 116].

Пермутационный подход, идея которого восходит к работам Р. Фишера [117], является одним из наиболее строгих непараметрических методов. Его преимущество заключается в том, что он является точным при фиксированном объеме набора данных и не требует предположений о виде распределений [113, 114, 118, 119]. Долгое время его широкое применение было ограничено высокой вычислительной сложностью, связанной с необходимостью перебора большого числа возможных перестановок между наблюдаемыми и предсказанными наборами данных. Однако современные вычислительные возможности позволяют эффективно реализовывать такие процедуры.

Основная идея пермутационных методов заключается в том, что выполняются все или заранее определенное число перестановок данных, для каждой из которых вычисляется некоторая выбранная исследователем из предметных соображений статистика, наблюдаемое распределение которой принимается за референтное распределение.

Пусть заданы наблюдаемые значения $C = C(\boldsymbol{s_1}), \dots, C(\boldsymbol{s_n})$ и предсказания модели $\hat{C} = \hat{C}(\boldsymbol{s_1}), \dots, \hat{C}(\boldsymbol{s_n})$. Определяется статистика интереса $T(C, \hat{C})$ – например, коэффициент корреляции.

Алгоритм пермутационного теста включает следующие этапы:

1. Вычисление наблюдаемой статистики $T_{\text{obs}} = T(C, \hat{C})$;

2. Формирование множества случайных перестановок данных в соответствии с выбранной нулевой гипотезой;

3. Вычисление статистики T_i для каждой перестановки;

4. Определение эмпирического распределения T_i и вычисление *p-value* как доли перестановок, где $| T_i | \geq | T_{\text{obs}} |$.

Метод бутстрэпа (bootstrap) [115, 116] используется для оценки доверительных интервалов и устойчивости метрик модели. Он основан на генерации большого числа псевдовыборок (бутстрэп-выборок) из исходного набора данных с возвращением. Для каждой такой бутстрэп-выборки пересчитывается статистика интереса, в результате формируется эмпирическое распределение интересующей статистики. Алгоритм bootstrap включает следующие этапы:

1. Вычисление наблюдаемой статистики $T_{\text{obs}} = T(C, \hat{C})$;

2. Генерация множества бутстрэп-выборок $\left(C_i^*, \hat{C}_i^*\right)$ путем случайного извлечения пар $\left(C(\boldsymbol{s_i}), \hat{C}(\boldsymbol{s_i})\right)$ из исходных данных с возвращением;

3. Вычисление статистики

$$T_i^* = T\left(C_i^*, \hat{C}_i^*\right)$$

для каждой бутстрэп-перестановки;

4. Формирование эмпирического распределения статистики $\{T_i^*\}$;

5. Оценка доверительных интервалов

$$\left[T_{(\alpha/2)}^*, T_{(1-\alpha/2)}^*\right],$$

или вычисление *p-value* как доли бутстрэп-оценок, удовлетворяющих $| T_i^* | \geq | T_{\text{obs}} |$.

В отличие от перестановочных тестов, бутстрэп не нарушает соответствие между наблюдаемыми и предсказанными значениями, оценивая вариабельность статистики за счет случайного повторного выбора наблюдений.

Метод *jackknife*, предложенный Куэнойлем [120] и развитый Тьюки [121], является вариантом ресэмплинга, при котором последовательно исключается по одному наблюдению из набора данных. Алгоритм jackknife включает следующие этапы:

1. Для каждого $i = 1, \ldots, n$ сформировать усеченную выборку $C_{(i)} = C \backslash \{C(\boldsymbol{s_i})\}$, $\hat{C}_{(i)} = \hat{C} \backslash \{\hat{C}(\boldsymbol{s_i})\}$,

2. Вычислить статистику для каждой усеченной выборки:

$$T_{(i)} = T\left(C_{(i)}, \hat{C}_{(i)}\right);$$

3. Найти среднее значение

$$\bar{T} = \frac{1}{n}\sum_{i=1}^{n} T_{(i)};$$

4. Оценить дисперсию (jackknife variance):

$$Var_{jack}(T) = \frac{n-1}{n}\sum_{i=1}^{n}\left(T_{(i)} - \bar{T}\right)^2;$$

5. При необходимости вычислить скорректированную оценку смещения (bias-corrected estimate):

$$T_{jack} = nT\left(C, \hat{C}\right) - (n-1)\bar{T}.$$

В отличие от бутстрэпа, который использует случайное извлечение с возвращением, jackknife систематически исключает по одному наблюдению, что делает его менее вычислительно затратным и более интерпретируемым при малых выборках. Метод jackknife используется для оценки дисперсии и смещения метрик качества, но реже применяется при анализе моделей, чем bootstrap.

1.4.5 Подходы к оценке устойчивости моделей

Еще одной задачей при построении моделей пространственного распределения примеси является оценка их устойчивости к изменениям входных

данных и структуры признакового пространства. Под устойчивостью модели в настоящей работе понимается чувствительность оценок коэффициентов регрессионной модели к вариациям исходных данных, а также к наличию зависимостей между предикторами.

При построении линейной модели решается задача подбора вектора коэффициентов модели, который лучше всего объясняет зависимость между наблюдениями и предикторами (1.4). Чтобы подобрать такой вектор коэффициентов, в классической линейной регрессии решают уравнение (1.5). Если столбцы $\mathbf{X}$ сильно коррелированы, то матрица $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ становится почти вырожденной. Численно это выражается в том, что матрица $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ становится плохо обусловленной, и ее обращение $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ становится нестабильным: малые изменения в данных приводят к большим ошибкам в $\widehat{\boldsymbol{\beta}}$.

Такое поведение характерно для мультиколлинеарности, то есть ситуации, когда один или несколько признаков можно выразить как (почти) линейную комбинацию других. Для пространственных переменных на основе вложенных круговых окрестностей мультиколлинеарность возникает естественно. Например, для точки наблюдения $\boldsymbol{s_i}$ пространственные переменные $v_j(\boldsymbol{s_i}, r)$ строятся на основе возрастающей последовательности вложенных областей радиуса r.

Отсюда следует, что для двух радиусов $r_1 < r_2$ выполняется соотношение:

$$v_j(\boldsymbol{s_i}, r_2) = v_j(\boldsymbol{s_i}, r_1) + \Delta v_j(\boldsymbol{s_i}; r_1, r_2),$$

где $\Delta v_j(\boldsymbol{s_i}; r_1, r_2)$ – вклад предиктора j в кольцевой области между радиусами r_1 и r_2.

Из этого соотношения следует наличие почти линейной зависимости между предикторами $v_j(\boldsymbol{s_i}, r_1)$ и $v_j(\boldsymbol{s_i}, r_2)$, формирующими соответствующие столбцы матрицы $\mathbf{X}$. Так в $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ появляются столбцы, которые почти повторяют друг друга. Это приводит, с одной стороны, к неустойчивости оценок $\widehat{\boldsymbol{\beta}}$: даже малые изменения в данных приводят к значительным колебаниям значений коэффициентов. С другой стороны, снижается интерпретируемость пространственных переменных: их вклад распределяется между несколькими вложенными окрестностями.

Для анализа мультиколлинеарности широко применяются методы, основанные на исследовании структуры признакового пространства. К числу базовых инструментов относится матрица парных корреляций, позволяющая выявить наличие линейной зависимости между отдельными предикторами. Однако парные корреляции не всегда позволяют выявить более сложные зависимости, возникающие при совместном влиянии нескольких переменных.

Более информативным показателем является фактор инфляции дисперсии (Variance Inflation Factor, VIF), который характеризует, насколько дисперсия оценки коэффициента регрессии увеличивается из-за линейной зависимости с другими предикторами [75, 122]. VIF определяется на основе коэффициента детерминации вспомогательной регрессии и позволяет выявлять скрытую мультиколлинеарность, не обнаруживаемую при анализе только парных корреляций. В прикладных задачах часто используются эмпирические пороговые значения VIF (например, 5 или 10), превышение которых свидетельствует о необходимости исключения или трансформации соответствующих признаков.

Другим подходом к оценке устойчивости является анализ обусловленности матрицы предикторов. В численном анализе число обусловленности функции измеряет, насколько может измениться выходное значение функции при небольшом изменении входного аргумента. При высокой обусловленности даже незначительные изменения в данных могут приводить к существенным изменениям оценок параметров модели. С точки зрения линейной алгебры число обусловленности определяется как отношение максимального и минимального сингулярных значений матрицы предикторов. Большие значения числа обусловленности указывают на близость матрицы к вырожденной и, как следствие, на потенциальную неустойчивость модели [122, 123].

Мультиколлинеарность и плохая обусловленность матрицы признаков связаны между собой. Наличие линейно зависимых или почти зависимых признаков приводит к уменьшению минимального сингулярного значения матрицы и росту числа обусловленности. В результате задача оценки параметров

становится плохо обусловленной, а получаемые оценки – чувствительными к шуму в данных.

1.4.6 Подходы к оценке репрезентативности сети наблюдений

Репрезентативность сети наблюдений является еще одним фактором, определяющим качество и устойчивость моделей пространственного распределения примесей. Способность модели описывать и прогнозировать значения на всей исследуемой территории определяется тем, насколько используемый для обучения модели набор точек наблюдений отражает пространственную структуру исследуемой территории. Под репрезентативностью в общем случае понимается степень, в которой наблюдения охватывают характерные условия формирования содержания примесей. В экологических и геоинформационных исследованиях подчеркивается, что при недостаточной репрезентативности даже формально корректно построенные модели могут давать смещенные и недостоверные оценки [124].

Традиционно набор данных считается репрезентативным, если он получен с использованием корректной схемы отбора (например, случайной или стратифицированной) и обеспечивает несмещенные оценки параметров генеральной совокупности. Показано, что вероятностные методы отбора обеспечивают более надежные оценки по сравнению с целенаправленным или «удобным» отбором, который может приводить к существенным смещениям результатов [125]. При этом репрезентативность часто не оценивается напрямую, а считается обеспеченной за счет выбранной схемы отбора [126]. Современные исследования подчеркивают, что на практике точки наблюдений часто размещаются исходя из логистических или административных ограничений, что приводит к смещенной и неконтролируемой репрезентативности [127].

В работах [124–127] репрезентативность рассматривается независимо от модели. В результате выборка может быть формально репрезентативной, но при этом давать неустойчивую или некачественную модель. В более современных исследованиях репрезентативность рассматривается как свойство данных,

определяемое их вкладом в качество и устойчивость модели. В машинном обучении такая концепция развивается в рамках задач оценки ценности данных (data valuation) и активного обучения, где наблюдения отбираются на основе их вклада в улучшение качества модели [128]. В этом подходе наблюдения рассматриваются как обладающие различной информативностью, а их вклад в обучение модели оценивается количественно. В частности, в работах [129] предлагается оценивать репрезентативность через изменение метрик качества модели при добавлении или исключении наблюдений. Аналогичные идеи лежат в основе методов анализа влияния наблюдений, где оценивается изменение параметров модели при варьировании обучающего подмножества [130]. В этом контексте отдельные наблюдения или их группы могут рассматриваться как информативные (повышающие качество модели), избыточные, или даже вредные (снижающие устойчивость и точность модели).

Таким образом, репрезентативность может быть определена через вклад наблюдений в качество модели. Такой подход позволяет перейти от формального анализа выборки к функциональной оценке данных и служит основой для разработки методов оптимального отбора наблюдений в задачах пространственного моделирования.

1.5 Выводы по главе 1

Таким образом, по результатам проведенного в главе обзора можно сделать следующие выводы:

1. Физические модели массопереноса обладают высокой интерпретируемостью, позволяют учесть химические превращения и метеофакторы, но требуют значительного объема входных данных и вычислительных ресурсов. Геостатистические методы опираются на предположение о пространственной автокорреляции и стационарности процессов, что ограничивает их применимость в условиях пространственно неоднородных сред. Модели машинного обучения и нейросетевые подходы способны выявлять сложные нелинейные зависимости и обеспечивать высокую точность прогноза,

однако их качество существенно зависит от числа, плотности, конфигурации и репрезентативности наблюдений. При малом числе наблюдений модели машинного обучения и нейросетевые модели сводятся, в сущности, к линейным моделям, не давая преимуществ перед классическими регрессионными методами. В задачах экологического мониторинга, особенно в условиях ограниченного числа наблюдений, требования к репрезентативности данных зачастую не выполняются, что приводит к неустойчивости моделей и снижению достоверности прогнозов.

2. На фоне указанных ограничений обоснована целесообразность использования подхода Land Use, который, помимо данных об измеренных содержаниях примесей в точках наблюдений, использует доступные географические данные для построения пространственного распределения примесей и позволяет ограничиться небольшим количеством измерений. В классическом Land Use подходе в качестве модели выступает линейная регрессия. В настоящей работе подход Land Use отделен от классической реализации с линейной регрессией (Land Use Regression) как самостоятельный этап построения предикторов на основе географических данных. Такой подход позволяет рассматривать Land Use в более общем смысле: как подход к построению предикторов, после которого может применяться любая модель аппроксимации – от линейной регрессии до нейронной сети.

3. Показано, что классическая реализация подхода Land Use имеет ряд ограничений. Использование вложенных круговых окрестностей точек наблюдений при формировании пространственных переменных приводит к многократному учету одних и тех же источников примеси, что вызывает мультиколлинеарность предикторов, ухудшает обусловленность матрицы предикторов и приводит к неустойчивости оценок параметров моделей. Кроме того, такая структура пространственных переменных затрудняет интерпретацию вклада источника в содержание примеси в точке наблюдения на различных расстояниях.

4. Анализ методов оценки качества моделей показал, что традиционно используемые метрики (абсолютная ошибка, среднеквадратичная ошибка,

коэффициент корреляции и др.) позволяют сравнивать конкурирующие модели между собой, однако оказываются малоинформативными в ситуации, когда рассматривается единственная модель. Применение асимптотических статистических тестов в условиях малых наборов данных может приводить к некорректным выводам о качестве модели. Это определяет необходимость использования непараметрических методов, в частности пермутационного подхода, позволяющего получать более надежные оценки значимости без строгих предположений о распределении данных.

5. Выявлено, что в опубликованных работах основное внимание уделяется построению и валидации моделей, в то время как влияние пространственного распределения и количества наблюдений на устойчивость и достоверность результатов рассматривается ограниченно. При этом в контексте Land Use моделирования репрезентативность сети наблюдений, как правило, трактуется в терминах схемы отбора и пространственного покрытия территории и редко анализируется с точки зрения вклада отдельных наблюдений или их групп в качество модели. Таким образом, недостаточно разработанным остается подход к оценке репрезентативности как вклада отдельных наблюдений или их групп в качество Land Use модели.

Глава 2. Материалы и методы

Исходные данные для диссертационного исследования были получены с помощью натурных измерений и синтетического моделирования.

2.1 Данные натурных измерений и выбор источников примесей

Использованы данные натурных наблюдений, полученные сотрудниками Института промышленной экологии УрО РАН в рамках ранее выполненных исследований (таблица 2.1). Отбор проб и химический анализ выполнялись по стандартным методикам, подробно описанным в работах [80, 131–133].

Таблица 2.1 Характеристики данных натурных наблюдений

Территория	Депонирующая среда	Число проб	Целевая переменная	Типы источников
Новый Уренгой	Снеговой покров	150	Поток пыли, мг/(м^2·сут)	основные и второстепенные дороги
Ноябрьск	Почва	200	Содержание металлов, мг/кг	–
Тарко-Сале	Почва	101	Содержание металлов, мг/кг	дороги, парковки
Мурманск	Поверхностная фация современных антропогенных отложений	40	Содержание марганца, мг/кг	шоссе, дороги

Источники примесей выбирались исходя из предметного предположения о наличии связи между расположением потенциальных источников и содержанием примеси в депонирующей среде. Состав и типология источников примесей различались для каждой исследуемой территории в зависимости от особенностей ее структуры (таблица 2.1). В настоящей работе используется идея, описанная в работе [60], согласно которой источники примесей рассматриваются как площадные источники.

2.2 Моделирование синтетической территории

Для проведения численных экспериментов была смоделирована синтетическая территория, представляющая собой квадратную область размером 5000×5000 м^2. Область дискретизировалась регулярной сеткой с шагом 50 м по обеим координатам. Таким образом, территория разбивалась на 100×100 ячеек, каждая из которых соответствовала элементарному прямоугольному сегменту.

На территории размещались источники площадного типа. Геометрия каждого источника – круг – задавалась с помощью центральной точки и радиуса, который варьировался в диапазоне от 50 до 200 м. Координаты точки привязки (центральная точка) источника выбирались равномерно случайно внутри границ моделируемой территории с учетом радиуса источника таким образом, чтобы любой источник радиуса $r \le r_{max}$ полностью помещался в границы территории:

$$c_x \sim U(x_{min} + r_{max}, x_{max} - r_{max}),$$
$$c_y \sim U(y_{min} + r_{max}, y_{max} - r_{max}),$$

где c_x, c_y – координаты центра источника,

$U(a, b)$ – случайное равномерное распределение на интервале $[a, b]$,

x_{min}, x_{max}, y_{min}, y_{max} – границы моделируемой территории,

r_{max} – максимальный радиус источника.

Плотность источников определялась как доля площади территории, занятой суммарно источниками. В настоящей работе суммарная площадь источников составляла около 10% территории, что позволяло моделировать умеренное покрытие.

Точки наблюдений размещались в пределах моделируемой территории равномерно случайно. Координаты каждой точки генерировались независимо в границах территории:

$$x_i \sim U(x_{\min}, x_{\max}), y_i \sim U(y_{\min}, y_{\max}).$$

При генерации учитывалось пространственное расположение источников примесей: точки наблюдения не допускались к размещению внутри областей, занятых источниками. Для этого каждая сгенерированная точка проверялась на пересечение с геометрией источников. В случае попадания внутрь хотя бы одного

источника точка отбрасывалась, и генерация повторялась. Процедура продолжалась до достижения заданного числа точек наблюдения. Всего генерировалось 100 точек наблюдений (таблица 2.2).

Таблица 2.2 Параметры синтетической территории

Параметр	Значение
Размер территории	5000 × 5000 м^2
Шаг сетки	50 м
Геометрия источников	круг
Радиусы источников	50–200 м
Доля площади источников (от площади территории)	0,1
Число точек наблюдений	100

2.3 Модели Land Use с предикторами на основе окрестностей точек наблюдений

Модели Land Use Buffers использовали пространственные предикторы, полученные с помощью классического подхода на основе вложенных круговых окрестностей точек наблюдений. Модели Land Use Rings предполагали построение пространственных переменных на основе непересекающихся кольцевых окрестностей точек наблюдений.

Структура каждой модели определялась с помощью компьютерного моделирования. Для каждой модели выполнялся подбор гиперпараметров через GridSearchCV (библиотека scikit-learn Python) с пространственной кросс-валидацией. Сетки гиперпараметров были выбраны исходя из размера набора данных (40–150 точек наблюдений) и числа предикторов (3–8), чтобы, с одной стороны, минимизировать риск переобучения и, с другой стороны, обеспечить разумное покрытие диапазонов значений гиперпараметров. Для ансамблевых методов и ИНС дополнительно контролировалась сложность моделей за счет ограничения глубины деревьев и числа параметров сети. Сетки гиперпараметров приведены в Приложении А.

2.3.1 Пространственные предикторы для моделей Land Use Buffers

Пусть имеется множество точек наблюдений $\boldsymbol{s_i} = (x_i, y_i)$, $i = 1, \ldots, n$, в которых выполнены измерения содержания примеси. Для каждой точки наблюдения $\boldsymbol{s_i}$ задавался дискретный набор радиусов $r_1, r_2, \ldots, r_K$, $r_1 < r_2 < \ldots < r_K$, определяющих систему вложенных окрестностей.

Круговая окрестность (буфер) радиуса r_k точки $\boldsymbol{s_i}$ определялась как множество

$$Buffer(\boldsymbol{s_i}, r_k) = \{\boldsymbol{s}: \|\boldsymbol{s} - \boldsymbol{s_i}\| \leq r_k\}, \qquad (2.1)$$

где $\|\cdot\|$ – евклидова метрика, которая при необходимости может быть заменена, например, на геодезическую метрику.

Таким образом, для точки $\boldsymbol{s_i}$ формировалась система вложенных множеств (рисунок 2.1)

$$Buffer(\boldsymbol{s_i}, r_1) \subseteq Buffer(\boldsymbol{s_i}, r_2) \subseteq \cdots \subseteq Buffer(\boldsymbol{s_i}, r_K). \qquad (2.2)$$

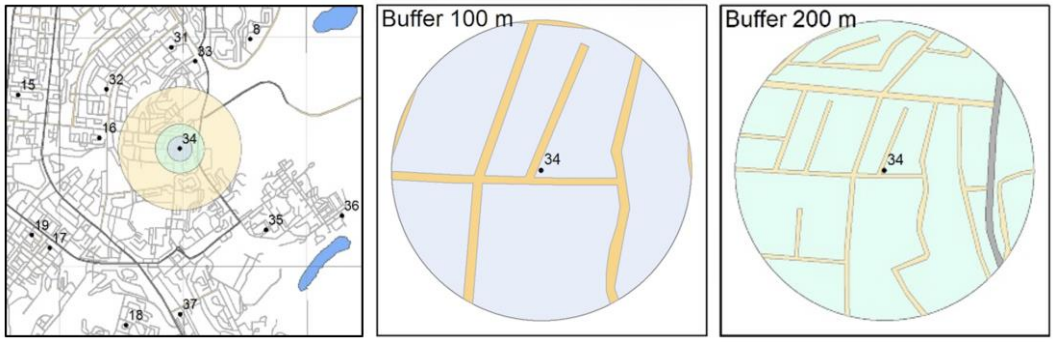


Рисунок 2.1 Подход к построению пространственных переменных Land Use Buffers

На следующем шаге выполнялось пересечение построенных круговых окрестностей с пространственными векторными и/или растровыми слоями данных, описывающими факторы, которые могут влиять на содержание примеси на территории.

Пусть территория описывается набором пространственных слоев $L = \{L_1, L_2, \ldots, L_T\}$, где каждый слой L_t представляет собой векторные или растровые данные, и каждый слой имеет свое геометрическое представление: пространственную геометрию для векторных слоев и атрибуты ячеек для растровых слоев.

Для вычисления пространственных переменных выполнялась операция пространственного пересечения, которая определяла часть слоя L_t, попадающую в круговую окрестность $Buffer(\boldsymbol{s_i}, r_k)$. В зависимости от типа геометрии слоя

количественная характеристика пространственной переменной определялась как агрегированная характеристика соответствующего пересечения *Buffer*($\boldsymbol{s_i}$, r_k) с пространственными слоями, вычисленная в пределах заданного радиуса r_k:

- для линейных объектов (например, дорог, линий электропередач и др.) – длина объектов внутри круговой окрестности:

$$Length(Buffer(\boldsymbol{s_i}, r_k) \cap L_t),$$

- для полигональных объектов (например, промышленных зон, жилых кварталов, лесные массивов) – площадь пересечения полигонов объектов и круговой окрестности:

$$Area(Buffer(\boldsymbol{s_i}, r_k) \cap L_t), \qquad (2.3)$$

- для точечных объектов (например, автозаправочных станций, вентиляционных шахт и др.) – количество объектов в круговой окрестности:

$$Count(Buffer(\boldsymbol{s_i}, r_k) \cap L_t),$$

- для растровых данных (например, плотность населения, доля покрытия территории типом землепользования) – сумма значений растра по ячейкам, попадающим в окрестность *Buffer*($\boldsymbol{s_i}$, r_k):

$$Sum(Buffer(\boldsymbol{s_i}, r_k) \cap L_t),$$

В настоящей работе рассматривались источники, представленные в виде полигональных объектов. Таким образом итоговый набор пространственных переменных для точки наблюдения $\boldsymbol{s_i}$ формировался путем агрегирования всех пространственных характеристик в пределах заданной окрестности r_k:

$$\boldsymbol{v_i} = \{v_i(r_1), v_i(r_2), \dots, v_i(r_K)\},$$

где $v_i(r_k)$ – значение пространственной переменной, представляющее собой агрегированную площадь по всем источникам примеси в окрестности радиуса r_k в точке наблюдения $\boldsymbol{s_i}$, $k = 1, \dots, K$ – индексы радиусов.

Тогда матрица пространственных переменных для всех n точек наблюдений определяется как

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{v}_1^{\mathrm{T}} \\ \boldsymbol{v}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{v}_n^{\mathrm{T}} \end{pmatrix}. \qquad (2.4)$$

Структура пространства вычисленных переменных приводит к нескольким ограничениям. Поскольку множества (2.2) являются возрастающей последовательностью, одно и то же полигональное множество может одновременно попадать в несколько окрестностей $Buffer(\boldsymbol{s_i}, r_k)$. В результате значения пространственных переменных

$$v_i(r_1), v_i(r_2), \ \dots, v_i(r_K)$$

частично дублируют друг друга, поскольку основаны на пересечении с одним и тем же набором объектов при различных радиусах.

Поскольку признаки $v_i(r_k)$ и $v_i(r_{k+1})$ часто оказываются линейно зависимыми, матрица предикторов **X** приобретает высокую коррелированность столбцов. Это затрудняет оценивание коэффициентов в линейных моделях и может приводить к нестабильности результатов.

2.3.2 Пространственные предикторы для моделей Land Use Rings

Подход к построению пространственных переменных на основе кольцевых окрестностей точек наблюдений представляет собой развитие классического Land Use подхода и позволяет дискретизировать пространственную область влияния источников примесей.

Пусть имеется множество точек наблюдений $\boldsymbol{s_i} = (x_i, y_i)$, $i = 1, \dots, n$, в которых выполнены измерения содержания примеси. Для каждой точки наблюдения $\boldsymbol{s_i} = (x_i, y_i)$ вводится система концентрических радиусов $0 = r_0 < r_1 < r_2 < \dots < r_K$.

На основе этих радиусов задаются кольцевые области:

$$Ring(\boldsymbol{s_i}, r_{k-1}, r_k) = \{\boldsymbol{s}: r_{k-1} \le \|\boldsymbol{s} - \boldsymbol{s_i}\| \le r_k\}, \qquad k = 1, \dots, K,$$

где $\|\cdot\|$ – евклидова метрика.

Каждая область $Ring(\boldsymbol{s_i}, r_{k-1}, r_k)$ представляет собой кольцевую окрестность точки наблюдения $\boldsymbol{s_i}$ (рисунок 2.2).

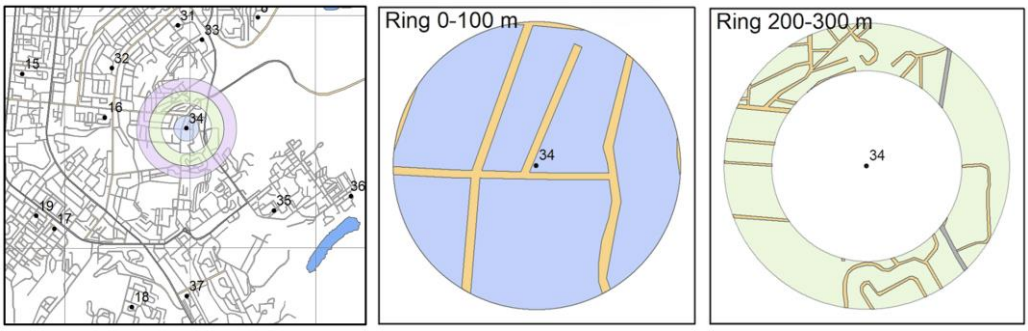


Рисунок 2.2 Подход к построению пространственных переменных Land Use Rings

При этом кольцевые окрестности попарно непересекающиеся:

$$Ring(\boldsymbol{s_i}, r_{k-1}, r_k) \cap Ring(\boldsymbol{s_i}, r_{l-1}, r_l) = \emptyset, k \neq l, \quad (2.5)$$

и выполняется условие

$$Ring(\boldsymbol{s_i}, r_{k-1}, r_k) = Buffer(\boldsymbol{s_i}, r_k) \backslash Buffer(\boldsymbol{s_i}, r_{k-1}).$$

Информация обо всех потенциальных типах источников примесей и их геометриях по-прежнему берется из пространственных слоев данных $L = \{L_1, L_2, \ldots, L_T\}$. Порядок вычисления пространственных переменных аналогичен изложенному в разделе 2.6.1, отличие заключается только в геометрии окрестностей точек наблюдений: вместо круга радиуса r_k используется кольцо, определяемое внутренним и внешним радиусом (r_{k-1}, r_k). Поскольку в настоящей работе учитываются источники, представленные в виде полигональных объектов, для каждой точки наблюдения $\boldsymbol{s_i}$ и каждого кольца $k = 1, \ldots, K$ вычислялась площадь пересечения с пространственными характеристиками:

$$v_i = \text{Area}(Ring(\boldsymbol{s_i}, r_{k-1}, r_k) \cap L_t). \quad (2.6)$$

Матрица пространственных переменных формировалась на основе величин площадей пересечения полигональных источников всех представленных типов и кольцевых окрестностей для n точек наблюдений

$$\boldsymbol{v}_i = \{v_i(r_0, r_1), v_i(r_1, r_2), \ldots, v_i(r_{K-1}, r_K)\},$$

где $r_0 = 0$, а $v_i(r_{k-1}, r_k)$ – агрегированная площадь пересечения источников с кольцевой окрестностью между радиусами r_{k-1} и r_k.

Набор радиусов для построения окрестностей точки наблюдения подбирался таким образом, чтобы отражать вероятное влияние источников примесей на исследуемой территории. В случае, когда нет четких данных о требуемых

величинах радиусов для характеристики тех или иных параметров, можно рассчитать большое количество переменных с различными размерами круговых окрестностей.

2.3.3 Модели аппроксимации Land Use Buffers и Land Use Rings

Для построения моделей аппроксимации Land Use Buffers и Land Use Rings использовались следующие архитектуры: линейная регрессия и ее регуляризованные модификации, ансамблевые алгоритмы (случайный лес, градиентный бустинг), метод опорных векторов, k-ближайших соседей и однослойный перцептрон.

Линейная регрессия. В настоящей работе использована множественная линейная регрессия и ее регуляризованные модификации. Стандартная линейная регрессия оценивала коэффициенты модели методом наименьших квадратов (Ordinary Least Squares, OLS) и выступала в качестве базовой (baseline) модели для сравнения с более сложными подходами.

При большом числе коррелированных предикторов, а также при ограниченном числе наблюдений, стандартная линейная регрессия может быть склонна к переобучению. Для решения этих проблем применялись методы регуляризации: Ridge-регрессия (L2-регуляризация), которая добавляет штраф на сумму квадратов коэффициентов, что ограничивает их величины и стабилизирует модель, и Lasso-регрессия (L1-регуляризация), которая использует штраф на сумму модулей коэффициентов, что может занулять незначимые коэффициенты и фактически осуществлять автоматический выбор предикторов.

Для Ridge коэффициент регуляризации α контролирует штрафную норму L2:

$$\min_{\beta}(\| \mathbf{C} - \mathbf{X}\boldsymbol{\beta} \|^2 + \alpha \| \boldsymbol{\beta} \|_2^2),$$

для Lasso коэффициент регуляризации α контролирует штрафную норму L1:

$$\min_{\beta}(\| \mathbf{C} - \mathbf{X}\boldsymbol{\beta} \|^2 + \alpha \| \boldsymbol{\beta} \|_1).$$

Малые значения коэффициента регуляризации выбирались для предотвращения чрезмерного зануления коэффициентов на небольших выборках.

Случайный лес. Случайный лес (Random Forest, RF) – ансамблевый алгоритм, основанный на построении множества независимых решающих деревьев. Каждое дерево обучается на случайном подмножестве обучающих данных с возвращением, и на каждом узле выбирается случайное подмножество предикторов. Итоговый прогноз формируется как среднее значение по деревьям ансамбля. RF устойчив к шуму, позволяет интерпретировать важность переменных.

Градиентный бустинг. Градиентный бустинг (например, такие реализации как XGBoost, LightGBM, CatBoost) часто превосходит RF по прогностической точности при удачной настройке. Градиентный бустинг представляет собой ансамблевый алгоритм, который строит последовательность деревьев решений, где каждое следующее дерево обучается на ошибках предыдущих. Итоговый прогноз формируется как взвешенная сумма прогнозов всех деревьев, где вес каждого дерева зависит от скорости обучения.

Метод опорных векторов. Метод опорных векторов (Support Vector Regression, SVR) ищет оптимальную гиперплоскость, аппроксимирующую зависимость между признаками и целевой переменной. В отличие от классических линейных моделей, SVR может использовать ядерные функции kernel, что позволяет моделировать нелинейные зависимости. Алгоритм устойчив к переобучению за счет параметра регуляризации C и допускает настройку точности аппроксимации через параметр ε.

Метод *k*-ближайших соседей. Метод k-ближайших соседей (k-Nearest Neighbors, kNN) – алгоритм регрессии, который предсказывает значение целевой переменной как усредненное значение ближайших соседей в пространстве признаков. В отличие от линейных моделей, kNN не строит явную зависимость между признаками и целевой переменной, а опирается на локальную структуру данных. kNN эффективен для небольших выборок и случаев, когда зависимость между признаками и целевой переменной сильно локализована или нелинейна.

Искусственные нейронные сети. В качестве ИНС использовалась модель на основе многослоного персептрона (MLP), которая, с одной стороны, продуктивна при прогнозировании нелинейных связей, и, с другой стороны, не требует

значительных вычислительных ресурсов. MLP относится к классу алгоритмов контролируемого обучения и представляет собой нейросетевую модель, способную аппроксимировать сложные нелинейные зависимости между входными признаками и целевой переменной. Для этого MLP использует один или несколько скрытых слоев с нелинейными функциями активации.

Входной слой MLP состоит из набора нейронов x_i, каждый из которых соответствует одному входному признаку. Нейрон скрытого слоя вычисляет взвешенную сумму входных сигналов:

$$z = w_1v_1 + w_2v_2 + \ldots + w_pv_p + b,$$

после чего к полученному значению применяется нелинейная функция активации.

Выходной слой получает значения из последнего скрытого слоя и преобразует их в итоговое предсказание модели. Для задач регрессии выходной нейрон обычно использует линейную функцию активации. Обучение MLP проводится методом обратного распространения ошибки с использованием градиентных методов оптимизации.

2.4 Модель Land Use Segmentation

Модель Land Use Segmentation основывалась на разбиении исследуемой территории на непересекающиеся множества (элементы), для каждого из которых оценивался вклад источников в формирование пространственного распределения примеси. В настоящей работе реализован частный случай разбиения территории, при котором элементы являлись прямоугольными сегментами, однако подход применим к произвольным выпуклым разбиениям.

В качестве исходных данных для моделирования использовалась величина потока примеси (мг/(м^2·сут)), которая рассчитывалась следующим образом

$$Flux = \frac{M}{S \cdot t},$$

где M – масса примеси (мг), которая определялась как произведение содержания (мг/л) на объем пробы (л), S – площадь отбора (м²), t – время экспозиции (сутки).

2.4.1 Прямая задача моделирования

Пусть Ω – исследуемая территория ($\Omega \subset \mathbb{R}^2$), $\Omega' \subset \Omega$ – область территории, занятая источниками. Функция вклада источника примеси моделируется скалярной функцией $f(\boldsymbol{a}, \boldsymbol{\theta})$, где $\boldsymbol{a} = \boldsymbol{a}(\boldsymbol{s}, \boldsymbol{s}')$ – вектор аргументов, зависящих от положений точек $\boldsymbol{s}$ и положений источников $\boldsymbol{s}'$, $\boldsymbol{\theta}$ – вектор параметров, характеризующий исследуемую территорию. Функция вклада $f(\boldsymbol{a}, \boldsymbol{\theta})$ определена на пространстве $\mathrm{A} \times \Theta$, где A – пространство аргументов, Θ – параметрическое пространство, определяемых следующим образом:

$$\mathrm{A} = \left\{\left(a_1, a_2, \dots, a_{dim_a}\right) \middle| \; a_k^{\min} \le a_k \le a_k^{\max}, k = 1, \dots, dim_a\right\}$$
$$= \left[a_1^{\min}, a_1^{\max}\right] \times \left[a_2^{\min}, a_2^{\max}\right] \times \dots \times \left[a_{dim_a}^{\min}, a_{dim_a}^{\max}\right],$$
$$\Theta = \left\{\left(\theta_1, \theta_2, \dots, \theta_{dim_\theta}\right) \middle| \; \theta_l^{\min} \le \theta_l \le \theta_l^{\max}, l = 1, \dots, dim_\theta\right\}$$
$$= \left[\theta_1^{\min}, \theta_1^{\max}\right] \times \left[\theta_2^{\min}, \theta_2^{\max}\right] \times \dots \times \left[\theta_{dim_\theta}^{\min}, \theta_{dim_\theta}^{\max}\right].$$

Тогда содержание примеси в точке $\boldsymbol{s}$ равно сумме вклада совокупности глобальных и локальных источников:

$$C(\boldsymbol{s}) = \varphi(\boldsymbol{s}) + \varphi'(\boldsymbol{s}),$$

где $\varphi(\boldsymbol{s})$ – функция координат, определяемая совокупностью глобальных источников, вообще говоря, зависящая от $\boldsymbol{s}$, $\varphi'(\boldsymbol{s})$ – вклад совокупности локальных источников, а именно

$$\varphi'(\boldsymbol{s}) = \int_{\Omega'} f(\boldsymbol{a}(\boldsymbol{s}, \boldsymbol{s}'), \boldsymbol{\theta}) d\boldsymbol{s}'$$

Таким образом для скалярного поля содержания примеси имеем

$$C(\boldsymbol{s}) = \varphi(\boldsymbol{s}) + \int_{\Omega'} f(\boldsymbol{a}(\boldsymbol{s}, \boldsymbol{s}'), \boldsymbol{\theta}) d\boldsymbol{s}'.$$

Для прямой задачи модель пространственного распределения примеси есть любое представление функции вклада в виде квадратично интегрируемой функции

$C(\boldsymbol{s})$. Таким образом, множество моделей есть гильбертово пространство $L_2(\Omega)$, а зависимость модели от множества параметров $\boldsymbol{\theta} \in \Theta$ есть отображение

$$\Theta \to L_2(\Omega),$$

которое сопоставляет каждому набору параметров $\boldsymbol{\theta} \in \Theta$ пространственное распределение $C(\boldsymbol{s})$, принадлежащее пространству квадратично интегрируемых функций $L_2(\Omega)$.

2.4.2 Обратная задача

Пусть имеются данные наблюдений в конечном числе точек $\boldsymbol{s_i}$, $i = 1, \ldots, n$. Обозначим наблюдаемое значение содержания примеси в i-й точке наблюдения $C^{obs}(\boldsymbol{s_i})$ как сумму истинного значения $C(\boldsymbol{s_i})$, вычисляемого с помощью функции $C(\boldsymbol{s})$, принятой в прямой задаче, и некоторой ошибки ε_i:

$$C^{obs}(\boldsymbol{s_i}) = C(\boldsymbol{s_i}) + \varepsilon_i.$$

Для заданного набора параметров $\boldsymbol{\theta}$ определим модельные (прогнозируемые) значения

$$\hat{C}(s_i, \boldsymbol{\theta}).$$

Введем векторные обозначения:

$$\widehat{\boldsymbol{C}}(\boldsymbol{s_i}, \boldsymbol{\theta}) = \left(\hat{C}(\boldsymbol{s_1}, \boldsymbol{\theta}), \hat{C}(\boldsymbol{s_2}, \boldsymbol{\theta}), \ldots, \hat{C}(\boldsymbol{s_n}, \boldsymbol{\theta})\right),$$

$$\boldsymbol{C^{obs}}(\boldsymbol{s_i}) = \left(C^{obs}(\boldsymbol{s_1}), C^{obs}(\boldsymbol{s_2}), \ldots, C^{obs}(\boldsymbol{s_n})\right),$$

Тогда обратная задача заключается в определении параметров $\widehat{\boldsymbol{\theta}}$ модели, которые минимизируют функцию потерь, определяемую как квадрат расхождения наблюдаемых и модельных значений:

$$Loss(\boldsymbol{\theta}) = \sum_{i=1}^{n} w_i \left\| C^{obs}(\boldsymbol{s_i}) - \hat{C}(\boldsymbol{s_i}, \boldsymbol{\theta}) \right\|^2,$$

где $w_i \geq 0$ — вес i-й точки наблюдения, характеризующий ее вклад в оценку параметров модели.

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}}\, Loss(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \left(C^{obs}(\boldsymbol{s_i}) - \hat{C}(\boldsymbol{s_i}, \boldsymbol{\theta}) \right)^2.$$

После определения параметров $\widehat{\boldsymbol{\theta}}$ модель используется для картирования смоделированного пространственного распределения содержания примеси на исследуемой территории.

2.4.3 Функция вклада

В общем случае вектор аргументов $\boldsymbol{a}(\boldsymbol{s}, \boldsymbol{s}')$ функции вклада может включать расстояние между точкой наблюдения и источником $d(\boldsymbol{s}, \boldsymbol{s}')$, угол между направлением на источник и направлением на север $\alpha(\boldsymbol{s}, \boldsymbol{s}')$, характеристики рельефа и подстилающей поверхности $z(\boldsymbol{s}, \boldsymbol{s}')$, а также, при необходимости, иные пространственные переменные:

$$\boldsymbol{a}(\boldsymbol{s}, \boldsymbol{s}') = (d(\boldsymbol{s}, \boldsymbol{s}'), \alpha(\boldsymbol{s}, \boldsymbol{s}'), z(\boldsymbol{s}, \boldsymbol{s}'), \dots).$$

Для определения вклада источников примеси использовалась физически мотивированная параметризованная функция, основанная на классической гауссовой модели рассеяния. В качестве исходной модели распределения примеси использовалось уравнение адвекции–диффузии (1.2) для пассивной примеси для стационарного случая с переносом вдоль оси *x*.

Для точечного источника уравнение (1.2) имеет известное решение в виде гауссова факела [20, 22], характеризующегося экспоненциальным убыванием содержания примеси в поперечном направлении и наличием характерных масштабов рассеяния:

$$C(x, y, z) = \frac{Q}{2\pi u \sigma_y(x) \sigma_z(x)} \exp\left(-\frac{y^2}{2\sigma_y^2(x)}\right)\left[\exp\left(-\frac{(z-H)^2}{2\sigma_z^2(x)}\right) + \exp\left(-\frac{(z+H)^2}{2\sigma_z^2(x)}\right)\right]$$

В настоящей работе рассматривались приземные источники, где $H \approx 0$ и оценивалось содержание примеси, интегрированное по оси z (масса примеси в столбе воздуха):

$$C(x, y) = \int_0^{+\infty} C(x, y, z)\, dz.$$

С учетом указанных допущений уравнение (2.10) принимает вид:

$$C(x, y) = \frac{Q}{\sqrt{2\pi}\, u\, \sigma_y(x)} \exp\left(-\frac{y^2}{2\sigma_y^2(x)}\right),$$

где $\sigma_y(x)$ – параметр поперечного рассеяния, который описывает распределение примеси в направлении, перпендикулярном переносу (оси *y*) на расстоянии xот источника.

Мы рассматриваем ситуацию с выделенным направлением оси x – вдоль факела рассеяния. Ось *Oz* выбирается как ортогональная к поверхности земли, а *Oy* – перпендикулярно к *Ox* (поперек факела). Рассмотрим фиксированную декартову систему координат с началом в источнике. В этой системе координат направление оси факела (оси *Ox*) задается ортом

$$\mathbf{a} = (\cos\varphi, \sin\varphi),$$

где φ – угол ориентации факела.

Тогда поперечная координата y, входящая в выражение для концентрации, может быть представлена как проекция радиус-вектора точки наблюдения $r = (x, y)$ на направление, ортогональное **a**.

При моделировании рассматривалось усреднение по множеству наблюдений, выполненных при различных направлениях ветра φ, которые предполагались равновероятными. В результате анизотропия, связанная с направлением переноса примеси, сглаживается:

$$\sigma_y(x) \approx \sigma.$$

Тогда функция вклада принимает радиально-симметричный вид:

$$C(d) = \frac{Q}{\sqrt{2\pi}\, u\, \sigma} \exp\left(-\frac{d^2}{2\sigma^2}\right), \qquad (2.7)$$

где $d = \sqrt{x^2 + y^2}$ – расстояние до источника.

Полученная функция (2.7) содержит параметры, характеризующие физические свойства источника и среды: мощность источника *Q*, скорость переноса *u* и масштаб рассеяния σ. В практических задачах экологического мониторинга эти параметры, как правило, неизвестны или определяются с существенной неопределенностью.

В связи с этим осуществляется переход к параметрической форме функции вклада, в которой указанные физические параметры заменяются эффективными параметрами, оцениваемыми по данным. В частности, множитель $\frac{Q}{\sqrt{2\pi}\, u\, \sigma}$

интерпретируется как вес вклада источника и заменяется параметром θ_1, а параметр σ, характеризующий пространственный масштаб рассеяния примеси, заменяется параметром θ_2.

Таким образом, функция вклада принимает вид:

$$f(d) = \theta_1 \exp\left(-\frac{d^2}{2\theta_2^2}\right) \quad (2.8)$$

2.4.4 Численная реализация. Дискретизация области источников

Для численной реализации алгоритма решения обратной задачи область источников Ω' разбивается на конечное число сегментов

$$\Omega' = \bigcup_{m=1}^{M} \Omega'_m.$$

Каждый сегмент Ω'_m характеризуется центроидом $\boldsymbol{s'_m} \in \Omega'$.

Тогда суммарное содержание примеси в точке $\boldsymbol{s}$ задается следующим образом:

$$C(\boldsymbol{s}) = \varphi(\boldsymbol{s}) + \sum_{m=1}^{M} f(d(\boldsymbol{s}, \boldsymbol{s'_m}), \boldsymbol{\theta}),$$

где $\varphi(\boldsymbol{s}) = \theta_0$ – принимается константой, предметно интерпретируемой как фоновое содержание примеси в любой точке исследуемой территории, $\theta_0^{\min} \leq \theta_0 \leq \theta_0^{\max}$.

Для пары источник-примесь вводится физически мотивированная параметризованная функция вклада $f(d, \boldsymbol{\theta})$, определенная на пространстве $\mathbb{R}_+ \times \Theta$:

$$f(d, \boldsymbol{\theta}) = \theta_1 \exp\left(-\frac{d^2}{2\theta_2^2}\right),$$

где $d(\boldsymbol{s}, \boldsymbol{s'}) = \|\boldsymbol{s} - \boldsymbol{s'}\|$ – евклидово расстояние от точки наблюдения $\boldsymbol{s}$ до точки области $\boldsymbol{s'}$, занятой источником, $\boldsymbol{\theta} = (\theta_1, \theta_2)$ – вектор параметров: θ_1 – вес вклада источника, θ_2 – характеристика масштаба рассеяния примеси.

Введем расстояние d_m от точки наблюдения $\boldsymbol{s}$ до центроида m-го сегмента $\boldsymbol{s'_m}$

$$d_m = \|\boldsymbol{s} - \boldsymbol{s'_m}\|,$$

отсюда

$$C(\boldsymbol{s}) = \theta_0 + \sum_{m=1}^{M} f(d_m, \boldsymbol{\theta}) = \theta_0 + \theta_1 \sum_{m=1}^{M} \exp\left(-\frac{\|\boldsymbol{s} - \boldsymbol{s'_m}\|^2}{2\theta_2^2}\right).$$

Для обратной задачи прогнозируемое содержание примеси в точке наблюдения $\boldsymbol{s_i}$, $i = 1, \ldots, n$, определяется следующим образом

$$\hat{C}(\boldsymbol{s_i}, \boldsymbol{\theta}) = \theta_0 + \theta_1 \sum_{m=1}^{M} \exp\left(-\frac{\|\boldsymbol{s_i} - \boldsymbol{s'_m}\|^2}{2\theta_2^2}\right).$$

Тогда решение обратной задачи сводится к определению параметров $\widehat{\boldsymbol{\theta}}$ модели, минимизирующих функцию потерь

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \left(C^{obs}(\boldsymbol{s_i}) - \theta_0 - \theta_1 \sum_{m=1}^{M} \exp\left(-\frac{\|\boldsymbol{s_i} - \boldsymbol{s'_m}\|^2}{2\theta_2^2}\right)\right)^2.$$

В настоящей работе использовались одинаковые веса наблюдений w_i.

2.5 Оценка устойчивости моделей Land Use Buffers и Land Use Rings

Следующий шаг после построения моделей с пространственными переменными на основе окрестностей точек наблюдений – отбор из вычисленных переменных наиболее информативных предикторов для выбранных моделей. Первоначальный набор включал все пространственные переменные, вычисленные для круговых и кольцевых окрестностей точек наблюдений. Далее выполнялись два вида анализа: (1) все пары переменных были проверены на мультиколлинеарность – наличие сильной линейной связи между переменными регрессионной модели, (2) оценивалась устойчивость моделей к выбору предикторов и изменениям в данных.

2.5.1 Анализ мультиколлинеарности

Для оценки линейных связей между пространственными переменными строилась матрица парных корреляций и вычислялся фактор инфляции дисперсии (Variance Inflation Factor – VIF) [75]. VIF оценивал, насколько дисперсия коэффициента регрессии завышена из-за мультиколлинеарности, и определялся как:

$$VIF_j = \frac{1}{1-R_j^2}, \quad (2.9)$$

где R^2 – нескорректированный коэффициент детерминации регрессии, в которой j-й предиктор играет роль отклика, а остальные предикторы – роль регрессоров. Чем выше VIF, тем выше вероятность наличия мультиколлинеарности – такая ситуация требует дальнейших исследований. VIF интерпретировался следующим образом:

- VIF = 1 – отсутствие линейной зависимости;

- VIF от 1 до 5 – умеренная зависимость;

- VIF > 5 – мультиколлинеарные предикторы.

В результате отбрасывались все предикторы со значением VIF, превышающим 5.

2.5.2 Оценка обусловленности матрицы предикторов

Для оценки устойчивости моделей вычислялось число обусловленности матрицы предикторов по L_2-норме:

$$\kappa(\mathbf{X}) = \frac{\sigma_{max}(\mathbf{X})}{\sigma_{min}(\mathbf{X})}, \quad (2.10)$$

где $\sigma_{max}(\mathbf{X})$ и $\sigma_{min}(\mathbf{X})$ это максимальное и минимальное сингулярные значения матрицы $\mathbf{X}$.

В настоящей работе число обусловленности матрицы пространственных переменных $\mathbf{X}$ вычислялось через собственные значения матрицы Грама $\mathbf{G} = \mathbf{X}^T\mathbf{X}$, и дополнительно с использованием функции numpy.linalg.cond библиотеки NumPy Python, которая вычисляет число обусловленности $\kappa(\mathbf{X})$ на основе полного сингулярного разложения (Singular Value Decomposition) матрицы $\mathbf{X}$.

Интерпретация значений $\kappa(\mathbf{X})$ была следующей:

- $\kappa(\mathbf{X}) \approx 1$ – матрица идеально обусловлена,

- $\kappa(\mathbf{X}) = 10–100$ – матрица допустимо (умеренно) обусловлена,

- $\kappa(\mathbf{X}) > 1000$ – матрица плохо обусловлена, модель потенциально численно неустойчива.

2.5.3 Стандартизация предикторов

Для оценки влияния масштаба предикторов на число обусловленности матрицы пространственных переменных стандартизировались через *z*-преобразование:

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \quad (2.11)$$

где x_{ij} – значение *j*-го пространственного предиктора в *i*-й точке наблюдения;

- $\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ – выборочное среднее по предиктору j;
- $\sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \mu_j)^2}$ – выборочное стандартное отклонение;
- $\tilde{x}_{ij}$ – стандартизированное значение предиктора.

2.6 Оценка качества моделей

2.6.1 Традиционные метрики

Пусть заданы два вектора значений одинаковой длины: $O = O(1), O(2), \ldots, O(n)$ и $P = P(1), P(2), \ldots, P(n)$, где O – наблюдаемые данные, P – предсказанные моделью данные, n – количество значений в наборе данных.

Точность прогноза оценивалась по следующим показателям: коэффициент детерминации R^2, коэффициент корреляции *Corr*, средняя абсолютная ошибка *MAE* и среднеквадратическая ошибка *RMSE*.

$$R^2 = 1 - \frac{\sum(P-\overline{P})^2}{\sum(O-\overline{O})^2} \quad (2.12)$$

$$Corr = \frac{\sum(P-\overline{P})(O-\overline{O})}{\sqrt{\sum(P-\overline{P})^2 \sum(O-\overline{O})^2}} \quad (2.13)$$

$$MAE = \frac{\sum|P-O|}{n} \quad (2.14)$$

$$RMSE = \sqrt{\frac{\sum(P-O)^2}{n}} \quad (2.15)$$

где O – наблюдаемый набор данных, P – прогнозируемый набор данных, $\overline{O}$ и $\overline{P}$ – средние значения наборов данных O и P, соответственно, n – размер набора данных (число наблюдений).

Полный перечень используемых метрик и их определения приведены в Приложении Б.

2.6.2 Пермутационный подход

В работе для оценки моделей предложено использовать пермутационный подход, который, в отличие от асимптотических тестов, не требует предположений о виде распределений наблюдаемых и предсказанных данных и является точным при конечном числе наблюдений [113, 114].

Выбор статистик. В настоящей работе использовались следующие статистики: разность средних $Diff_{means}$ (2.16), разность медиан $Diff_{medians}$ (2.17), коэффициент корреляции Пирсона $Corr(O, P)$ (2.18):

$$Diff_{means}(O,P) = \bar{P} - \bar{O} = \frac{\sum_{j=1}^{n} P(j)}{n} - \frac{\sum_{j=1}^{n} O(j)}{n} = \frac{\sum_{j=1}^{n} (P(j) - O(j))}{n}, \quad (2.16)$$

$$Diff_{medians}(O,P) = median(P) - median(O), \quad (2.17)$$

$$Corr(O,P) = \frac{\sum_{j=1}^{n} (O(j) - \bar{O})(P(j) - \bar{P})}{\sqrt{\sum_{j=1}^{n} (O(j) - \bar{O})^2 \sum_{j=1}^{n} (P(j) - \bar{P})^2}}, \quad (2.18)$$

где O – вектор наблюдаемых данных, P – вектор предсказанных данных, $\bar{O}, \bar{P}$ – средние значения наборов данных O и P, j – номер позиции в наборе данных, $j = 1, \dots n$, $median(O)$, $median(Y)$ – медианы наборов данных O и P.

Выбор статистик обусловлен тем, что они характеризуют различные аспекты согласия между наблюдаемыми и прогнозируемыми значениями. Разность средних и разность медиан описывают разность выраженности признака между двумя наборами данных. Коэффициент корреляции Пирсона, в свою очередь, характеризует степень согласованности паттернов («рисунков изменчивости») наблюдаемых и предсказанных значений с точностью до масштаба и сдвига.

Схема перестановок. Эмпирическое распределение выбранной статистики строилось путем многократного применения допустимых перестановок в зависимости от предполагаемой нулевой гипотезы и вычисления значения статистики на каждой перестановке. Затем вычислялся пермутационный *p-value* как доля перестановок, при которых значение статистики оказывалось равным или

более экстремальным, чем наблюдаемое значение при справедливости нулевой гипотезы. Схема перестановок определялась тестируемой нулевой гипотезой и выбранной статистикой (таблица 2.3).

Таблица 2.3 Статистики, использованные для пермутационного подхода

Статистика	Гипотеза H_0	Схема перестановок	Тип *p-value*
Разность средних значений	$F_O = F_P$, где F_O и F_P – распределения O и P соответственно	полные перестановки (перестановки объединенного набора данных)	Одностороннее *p-value* определялось как минимальное из двух односторонних (слева или справа)
Разность медиан			
Коэффициент корреляции	$Corr(O, P) = 0$	перестановки набора данных O при фиксированном P	Правостороннее *p-value*

Для статистик, зависящих только от одномерных распределений – таких как разность средних и медиан – проверялась гипотеза однородности распределений O и P. В этом случае применялись полные перестановки объединенного набора данных с последующим разбиением на две группы равного размера, так как если O и P происходят из одного и того же распределения, то принадлежность конкретного значения к группе O или P не имеет значения. Следовательно, все разбиения объединенного набора данных на две равные части равновероятны при справедливости H_0.

Для коэффициента корреляции Пирсона проверялась гипотеза об отсутствии связи между O и P в смысле выбранного коэффициента корреляции. В этом случае использовалась схема перестановок, при которой значения O случайным образом перемешивались при фиксированном P, поскольку при справедливости H_0 совместное распределение пар $(O(j), P(j))$ инвариантно к перестановке значений O, поскольку при отсутствии зависимости между переменными любая их случайная переассоциация эквивалентна исходной.

Пример. В таблице 2.4 приведен пример перестановок для выбранных статистик на примере набора данных длиной $n = 4$. В строке с индексом 0 (жирный шрифт) рассчитано наблюдаемое значение каждой из статистик.

Таблица 2.4 Пример перестановок для выбранных статистик

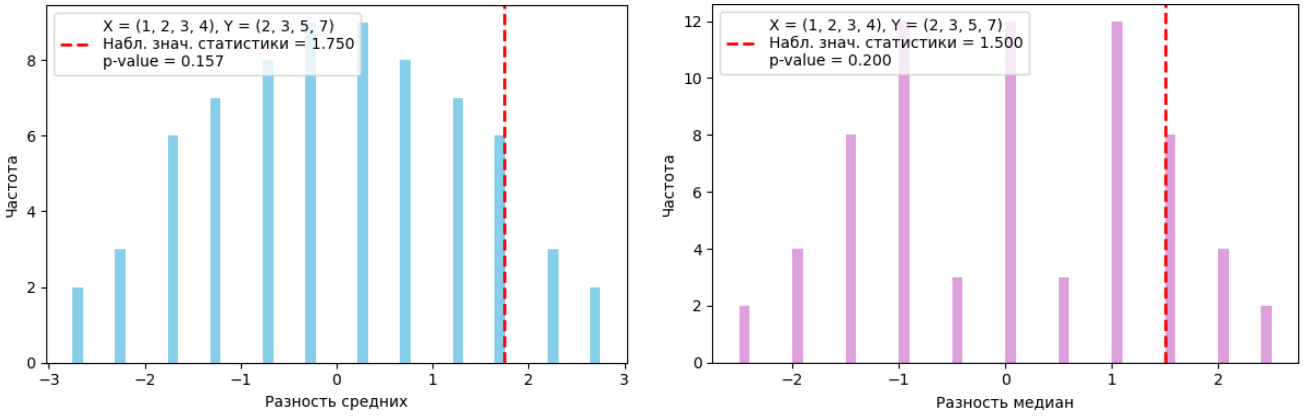
№	Перестановка Z	O'	P'	$Diff_{means}$	$Diff_{medians}$	№	O'	P	$Corr(O', P)$
0	**(1, 2, 3, 4, 2, 3, 5, 7)**	**(1, 2, 3, 4)**	**(2, 3, 5, 7)**	**1,75**	**1,50**	**0**	**(1, 2, 3, 4)**	**(2, 3, 5, 7)**	**0,9897**
1	(2, 3, 1, 5, 2, 4, 3, 7)	(2, 3, 1, 5)	(2, 4, 3, 7)	1,25	1,00	1	(1, 2, 4, 3)	(2, 3, 5, 7)	0,7568
2	(7, 5, 3, 2, 1, 2, 3, 4)	(7, 5, 3, 2)	(1, 2, 3, 4)	-1,75	-1,50	2	(1, 3, 2, 4)	(2, 3, 5, 7)	0,7568
3	(1, 5, 2, 7, 2, 3, 3, 4)	(1, 5, 2, 7)	(2, 3, 3, 4)	-0,75	-0,50	3	(1, 3, 4, 2)	(2, 3, 5, 7)	0,2911
…	…	…	…	…	…	…	…	…	…
69	(3, 2, 5, 4, 1, 2, 3, 7)	(3, 2, 5, 4)	(1, 2, 3, 7)	-0,25	-1,00	23	(4, 3, 2, 1)	(2, 3, 5, 7)	-0,9897

Исходные наборы данных $O = (1, 2, 3, 4)$ и $P = (2, 3, 5, 7)$, в этом случае объединенный набор данных $Z = (1, 2, 3, 4, 2, 3, 5, 7)$. Для коэффициента корреляции всего было возможно $4! = 24$ перестановки набора данных O. Для разности средних и разности медиан выполнялось

$$\binom{8}{4} = \frac{8!}{4!\,4!} = \frac{1680}{24} = 70,$$

возможных разбиений объединенного набора данных Z' на две группы O' и P' по 4 значения с учетом уникальности каждого наблюдения.

Для всех перестановок рассчитаны выбранные статистики; для каждой статистики построены гистограммы распределений. На каждой гистограмме добавлена вертикальная линия, показывающая наблюдаемое значение выбранной статистики (рисунок 2.3). Для получения *p-value* рассчитывалась доля перестановок, для которых значение статистики оказалось равным или более экстремальным, чем наблюдаемое значение.



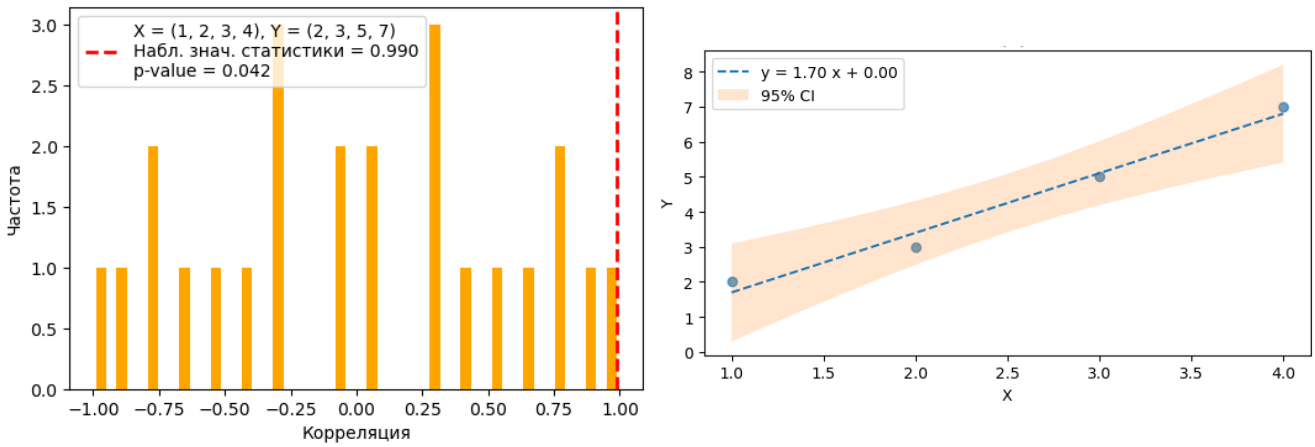


Рисунок 2.3 Пример применения пермутационного подхода для набора данных длиной $n = 4$

2.6.3 Диаграмма Тейлора

Для визуализации прогностической способности моделей использовалась диаграмма Тейлора, которая объединяет стандартное отклонение, коэффициент корреляции Пирсона и центрированную квадратичную ошибку (centered *RMS* difference) построенных моделей [109, 110]. Диаграмма Тейлора позволяет оценить, насколько точно предложенные методы прогноза моделируют наблюдаемое поле.

Пусть задано наблюдаемое скалярное поле

$$O(i): i = 1, \dots, n;$$

и предсказанное моделью скалярное поле

$$P(i): i = 1, \dots, n;$$

с весами

$$w(i): i = 1, \dots, n; \quad \sum w(i) = 1.$$

Тогда средние значения и дисперсии наблюдаемого и предсказанного полей определяются как

$$\bar{O} = \sum O(i)w(i)\,;\ \sigma_O^2 = \sum (O(i) - \bar{O})^2 w(i).$$

$$\bar{P} = \sum P(i)w(i)\,;\ \sigma_P^2 = \sum (P(i) - \bar{P})^2 w(i).$$

Корреляция наблюдаемого и моделируемого полей определяется как

$$Corr(O,P) = \frac{\sum(P(i)-\bar{P})(O(i)-\bar{O})w(i)}{\sqrt{\sum(P(i)-\bar{P})^2 w(i)}\sqrt{\sum(O(i)-\bar{O})^2 w(i)}} = \frac{\sum(P(i)-\bar{P})(O(i)-\bar{O})w(i)}{\sigma_P \sigma_O}.$$

Среднеквадратичная ошибка *RMSE*(*O*, *P*) между полями вычисляется как

$$RMSE(O,P) = \sqrt{\sum\big(P(i)-O(i)\big)^2 w(i)}, \qquad \sum w(i) = 1.$$

Выделим общее смещение:

$$\bar{E} = \bar{P} - \bar{O}$$

и центрированную квадратичную ошибку (*RMSE* центрированного паттерна, centered pattern *RMS* difference):

$$E' = \sqrt{\sum\big((P(i)-\bar{P})-(O(i)-\bar{O})\big)^2 w(i)} = \sqrt{\sigma_P^2 + \sigma_O^2 - 2\sigma_P\sigma_O Corr(O,P)}.$$

При этом *RMSE*(*O*, *P*) раскладывается на две компоненты:

$$RMSE^2(O,P) = E'^2 + \bar{E}^2.$$

Геометрическая интерпретация (через закон косинусов).

Центрированная *RMSE* связывает статистики следующим соотношением:

$$E'^2 = \sigma_P^2 + \sigma_O^2 - 2\sigma_P\sigma_O Corr(O,P),$$

где

$$cos(\varphi_P) = Corr(O,P),$$

$$\varphi_P = \cos^{-1}\big(Corr(O,P)\big).$$

Это полностью соответствует закону косинусов:

$$c^2 = a^2 + b^2 - 2ab\cos\varphi.$$

Наблюдаемое поле размещается на оси абсцисс диаграммы Тейлора на расстоянии σ_O от начала координат. Для предсказанного поля координаты определяются как

$$x_P = \sigma_P \cos\varphi = \sigma_P Corr(O,P)$$

$$y_P = \sigma_P \sin\varphi = \sigma_P \sin\big(\cos^{-1}\big(Corr(O,P)\big)\big)$$

Для максимального радиуса x_{max} координаты подписей дуговой шкалы корреляции задаются аналогично:

$$x = x_{max} \cos \varphi = x_{max} Corr(O, P)$$

$$y = x_{max} \sin \varphi = x_{max} \sin(\cos^{-1}(Corr(O, P)))$$

x_{max} – правая граница оси X.

На рисунке 2.4 приведен пример диаграммы Тейлора, построенной в полярной системе координат. Радиальные расстояния от начала координат представляют собой нормализованные стандартные отклонения, а азимутальные положения – коэффициенты корреляции между измеренными и прогнозируемыми распределениями. Концентрические круги представляют собой среднеквадратичное отклонение *RMSE*.

На диаграмме представлена эталонная модель, расположенная на оси стандартного отклонения. Чем ближе положение модели к эталонной точке, тем более точно воспроизводятся наблюдаемые данные.

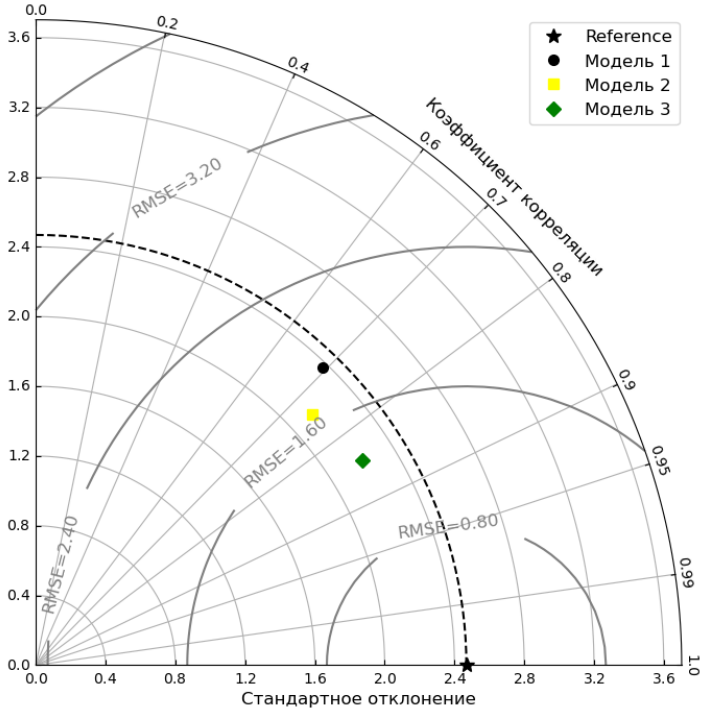


Рисунок 2.4 Пример построения диаграммы Тейлора для нескольких моделей

2.7 Метод оценки репрезентативности точек наблюдений

Для оценки репрезентативности точек наблюдений был реализован двухэтапный подход, основанный на многократном случайном разбиении набора данных и анализе вклада точки (группы точек) в качество модели.

Этап 1. Выбор архитектуры модели

На первом этапе осуществлялся подбор гиперпараметров модели на основе многослойного персептрона (MLP). На вход модели подавались пространственные координаты точек наблюдений, предварительно стандартизованные с использованием линейного масштабирования (StandardScaler).

Подбор гиперпараметров (размер скрытого слоя и коэффициент регуляризации) выполнялся с помощью процедуры GridSearchCV с 5-кратной кросс-валидацией. В качестве критерия качества использовалась среднеквадратическая ошибка *RMSE*. В результате подбиралась архитектура модели, которая затем использовалась на всех итерациях моделирования.

Этап 2. Оценка репрезентативности

Для оценки репрезентативности использовалась процедура многократного случайного разбиения исходного набора данных:

- исходный набор из *n* точек многократно случайным образом делился на обучающее (75%) и тестовое (25%) подмножества;

- для каждого разбиения обучалась модель с фиксированными на этапе 1 гиперпараметрами;

- на тестовом подмножестве рассчитывались метрики качества *RMSE* и *Corr* между наблюдаемыми и предсказанными значениями;

- для каждого разбиения сохранялся состав обучающего подмножества и соответствующие значения метрик.

Из всех полученных разбиений отбиралось подмножество «лучших» моделей, для которых значение *RMSE* находилось в нижнем дециле распределения (10% наименьших значений *RMSE*).

Наконец, для каждой *i*-й точки (группы точек) рассчитывалась метрика репрезентативности как среднее значение ошибки моделей, в обучающие подмножества которых входила данная точка:

$$R(\boldsymbol{s_i}) = \frac{1}{\mid K_i \mid \sum_{k \in K_i} RMSE_k},$$

где $K_i = \{\, k \mid \boldsymbol{s_i} \in Train_k \,\}$ – множество разбиений, для которых получены наилучшие значения *RMSE*, $\mid K_i \mid$ – число элементов (мощность) множества K_i.

2.8 Выводы по главе 2

1. Предложен подход к моделированию синтетической территории, предназначенный для контроля расчетных условий, валидации, тестирования устойчивости и качества моделей.

2. Показано, что классический подход к построению пространственных переменных-предикторов Land Use Buffers, при котором пространственные переменные вычисляются как площади пересечения концентрических круговых окрестностей точек наблюдений с предполагаемыми источниками примесей, приводит к многократному учету одних и тех же источников примеси. Это, в свою очередь, приводит к мультиколлинеарности пространственных предикторов модели и затрудняет интерпретацию вклада отдельных источников.

3. Предложен подход к построению пространственных переменных-предикторов Land Use Rings, при котором пространственные переменные вычисляются как площади пересечения непересекающихся кольцевых окрестностей точек наблюдений с предполагаемыми источниками примесей. Показано, что такой подход снижает мультиколлинеарность предикторов модели и повышает интерпретируемость вклада источников на разных расстояниях.

4. Предложено развитие Land Use подхода – переход от пространственных переменных, формируемых на основе окрестностей точек наблюдений, к покрытию исследуемой территории непересекающимися множествами (элементами). Для каждого элемента покрытия определяется набор типов источников примеси и функции, описывающие его вклад в формирование пространственного распределения примеси на исследуемой территории. При этом функция вклада может быть задана любым подходящим способом (например, аналитическим выражением или нейросетевой моделью, в том числе физически информированной). В настоящей работе реализован частный случай предложенного подхода – разбиение территории на прямоугольные сегменты (Land Use Segmentation). Land Use Segmentation позволяет решать прямую и обратную задачи: построение пространственного распределения примеси по заданной

функции вклада и восстановление функции вклада и построение пространственного распределения примеси по измерениям в конечном числе точек наблюдений.

5. Разработана физически информированная модель пространственного распределения примесей на основе подхода Land Use Segmentation, в которой используется параметризованная функция вклада гауссовского типа для построения пространственного распределения примесей при ограниченных данных наблюдений.

6. Построены три типа Land Use моделей: Land Use Buffers, Land Use Rings, Land Use Segmentation. Для оценки устойчивости моделей предложено использовать анализ мультиколлинеарности предикторов и числа обусловленности матрицы признаков. Качество моделей предложено оценивать с использованием совокупности подходов, включающих традиционные метрики качества, диаграмму Тейлора и предложенный пермутационный подход.

7. Разработаны принципы применения пермутационного подхода к оценке качества моделей пространственного распределения примесей. Предложенный подход не требует предположений о виде распределений и является точным при фиксированном размере набора данных. Предложена схема применения пермутационного подхода для трех статистик: разности средних значений, разности медиан и коэффициента корреляции Пирсона, однако, не ограничиваясь ими (выбор статистик остается за исследователем).

8. Предложен метод оценки индивидуальной и коллективной репрезентативности точек наблюдений для задач пространственного моделирования. Репрезентативность определяется как среднее значение ошибки моделей, в обучающие подмножества которых входила данная точка.

Все приведенные в главе методы и модели реализованы в виде комплекса программ для моделирования пространственного распределения примесей.

Результаты второй главы опубликованы в работах [8, 11, 12, 18], соответствующий комплекс компьютерных программ: [13].

Глава 3. Результаты и обсуждение

На основании разработанных численных методов и моделей был реализован программный комплекс на языке программирования Python 3.12, среда разработки – Visual Studio Code (VS Code). Исходный код программного комплекса доступен в открытом репозитории по ссылке https://github.com/amoskalyova/land-use.

Использовались следующие стандартные и специализированные библиотеки Python: math и itertools для базовых математических операций и перебора параметров, NumPy и SciPy для продвинутых математических операций и линейной алгебры, Pandas для работы с табличными и GeoPandas для работы с геопространственными данными, Shapely для геометрических операций, scikit-learn (sklearn) для машинного обучения, Matplotlib и Seaborn для визуализаций, rasterio для чтения, записи и обработки растровых геоданных.

Архитектура программного комплекса приведена на рисунке 3.1 и включает модули со справочной информацией о примесях, их источниках и функциях вклада источников примесей (lu_info.py), библиотеку функций для геометрических и численных операций на исследуемой территории (lu_lib.py), модули моделей Land Use Buffers/Rings (lu_ml.py) и Land Use Segmentation (segmentation.py), модули оценки устойчивости (robustness.py) и качества моделей (performance.py) и модуль оценки репрезентативности точек наблюдений (representativeness.py). Программный комплекс поддерживает запуск численных экспериментов на основе синтетических данных и данных наблюдений и расширение набора моделей/функций вклада, примесей, источников, исследуемых территорий.

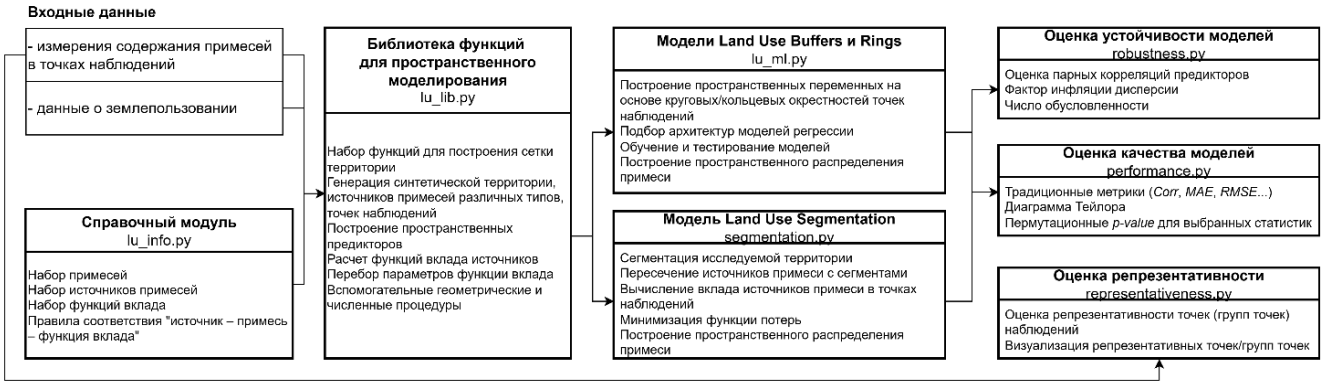


Рисунок 3.1 Архитектура программного комплекса для моделирования пространственного распределения примесей

3.1 Оценка устойчивости моделей

Раздел посвящен анализу свойств матрицы **X** пространственных переменных, построенных на основе круговых окрестностей точек наблюдений (2.3.1) и на основе непересекающихся кольцевых окрестностей точек наблюдений (2.3.2).

3.1.1 Постановка численного эксперимента

Для оценки устойчивости моделей проведен численный эксперимент на синтетической территории 5000×5000 м^2, описанной в разделе 2.2. Всего выполнено 100 независимых симуляций, в каждой из которых:

- генерировалось пространственное распределение источников примеси;
- генерировалось множество n = 100 точек наблюдений;
- для каждой точки наблюдения $\boldsymbol{s_i}$ формировались системы пространственных окрестностей с использованием единой конфигурации радиусов:

$$r_0 < r_1 < r_2 < \cdots < r_K, \qquad r_0 = 0,$$

где r_K = 1000 м, $r_k = k \cdot \Delta r$, Δr = 100 м.

На основе этой конфигурации для каждой точки наблюдения $\boldsymbol{s_i}$ определялись K = 10 круговых окрестностей

$$Buffer_i(100), Buffer_i(200), \dots, Buffer_i(1000),$$

и K = 10 смежных кольцевых окрестностей

$$Ring_i(0, 100), Ring_i(100, 200), \dots, Ring_i(900, 1000).$$

При такой конфигурации для любой точки наблюдения $\boldsymbol{s_i}$ первая кольцевая окрестность $Ring_i(0, 100)$ совпадала с первой круговой окрестностью $Buffer_i(100)$ и объединение кольцевых окрестностей эквивалентно максимальной круговой:

$$\bigcup_{k=1}^{K} Ring_i\,(r_{k-1}, r_k) = \bigcup_{k=1}^{K} Buffer_i\,(r_k).$$

Такое построение обеспечивало сопоставимость двух подходов к построению пространственных предикторов: в обоих случаях использовалась одна и та же область влияния радиуса r_K, но различался способ агрегации вклада источников (накопительный для круговых окрестностей и раздельный для кольцевых).

Затем для каждой точки наблюдения $\boldsymbol{s_i}$ и каждой окрестности вычислялись значения пространственных предикторов как площади пересечения источников примеси с соответствующей окрестностью. На основе этих значений формировались матрицы пространственных переменных-предикторов **X**.

Для каждой симуляции и каждого типа пространственных переменных-предикторов анализировалась структура матрицы **X**: рассчитывались парные корреляции между предикторами, оценивался фактор инфляции дисперсии и вычислялось число обусловленности для исходной и стандартизированной матрицы (см. раздел 2.5.1).

3.1.2 Результаты численных экспериментов на синтетических данных

Матрицы корреляций, усредненные для 100 симуляций, приведены в таблицах 3.1–3.2. Для пространственных переменных на основе круговых окрестностей точек наблюдений наблюдалась сильная линейная зависимость: значения коэффициентов корреляции между соседними по радиусу переменными достигали 0,86–0,97 и даже для существенно различающихся радиусов (например, r_k = 100 и r_k = 1000) оставались не ниже 0,19.

Таблица 3.1 Матрица корреляций для пространственных переменных на основе круговых окрестностей точек наблюдений

r_k	100	200	300	400	500	600	700	800	900	1000
100	1,00	0,86	0,67	0,50	0,38	0,32	0,27	0,24	0,21	0,19
200		1,00	0,91	0,73	0,56	0,45	0,38	0,33	0,29	0,25
300			1,00	0,91	0,75	0,61	0,51	0,44	0,38	0,33
400				1,00	0,93	0,79	0,66	0,57	0,50	0,44
500					1,00	0,94	0,83	0,72	0,63	0,57
600						1,00	0,95	0,85	0,76	0,68
700							1,00	0,96	0,87	0,79
800								1,00	0,97	0,89
900									1,00	0,97
1000										1,00

Пространственные переменные, формируемые на основе непересекающихся кольцевых окрестностей точек наблюдений, показали значительно более низкую

корреляцию между предикторами. Умеренные коэффициенты корреляции (от 0,65 до 0,77) наблюдались преимущественно между соседними кольцевыми окрестностями. При этом для несмежных кольцевых окрестностей коэффициенты корреляции были близки к нулю и в ряде случаев принимали отрицательные значения, что указывает на отсутствие линейной связи между соответствующими пространственными переменными.

Таблица 3.2 Матрица корреляций для пространственных переменных на основе кольцевых окрестностей точек наблюдений

(r_{k-1}, r_k)	(0, 100)	(100, 200)	(200, 300)	(300, 400)	(400, 500)	(500, 600)	(600, 700)	(700, 800)	(800, 900)	(900, 1000)
(0, 100)	1,00	0,77	0,42	0,10	–0,01	0,02	0,03	0,02	0,02	0,02
(100, 200)		1,00	0,75	0,28	–0,01	–0,01	0,02	0,02	0,00	–0,01
(200, 300)			1,00	0,68	0,20	0,01	–0,01	0,01	0,00	–0,01
(300, 400)				1,00	0,67	0,23	0,03	0,02	0,04	0,05
(400, 500)					1,00	0,70	0,27	0,08	0,07	0,11
(500, 600)						1,00	0,70	0,30	0,11	0,11
(600, 700)							1,00	0,72	0,32	0,13
(700, 800)								1,00	0,72	0,31
(800, 900)									1,00	0,72
(900, 1000)										1,00

Результаты расчета VIF также показали существенные различия между двумя типами пространственных предикторов (таблица 3.3).

Таблица 3.3 Оценки VIF

Переменные-предикторы	min	mean	median	max
На основе круговых окрестностей	6,64	136,36	129,42	284,22
На основе кольцевых окрестностей	2,90	7,06	7,31	10,57

Для предикторов, построенных на основе круговых окрестностей точек наблюдений, значения VIF характеризовались крайне высокой вариабельностью и достигали нескольких сотен (max = 284,22). Среднее значение VIF для предикторов на основе круговых окрестностей составило 136,36. Такие значения свидетельствуют о сильной мультиколлинеарности между переменными, обусловленной вложенной природой круговых окрестностей.

В противоположность этому, для предикторов, сформированных на основе кольцевых окрестностей, значения VIF оказались существенно ниже и значительно более устойчивыми: минимальное значение VIF составило 2,90, среднее – 7,06, медианное – 7,31, а максимальное значение не превышало 10,57. Эти значения соответствуют слабой или умеренной степени мультиколлинеарности [75].

Значения числа обусловленности $\kappa(\mathbf{X})$ вычислялись для исходных и стандартизированных матриц предикторов (таблица 3.4).

Таблица 3.4 Оценки $\kappa(\mathbf{X})$

Переменные-предикторы	min	mean	median	max
На основе круговых окрестностей (исходные)	305,72	407,27	398,82	554,69
На основе круговых окрестностей (стандартизированные)	48,78	70,03	68,21	104,56
На основе кольцевых окрестностей (исходные)	43,40	63,18	62,23	89,86
На основе кольцевых окрестностей (стандартизированные)	7,32	9,33	9,21	13,15

Уже на исходных данных переход к кольцевым окрестностям весомо снизил число обусловленности матрицы признаков по сравнению с круговыми окрестностями (63,18 против 407,27). После стандартизации переменных число обусловленности снижалось для обеих матриц, однако различия между круговыми и кольцевыми подходами сохранялись. Для круговых переменных $\kappa(\mathbf{X})$ уменьшалось примерно в 3–4 раза, оставаясь на порядок выше, чем у кольцевых. Для кольцевых переменных стандартизация снижала $\kappa(\mathbf{X})$ до в среднем 9,33.

Предложенный подход Land Use Rings к построению пространственных переменных на основе непересекающихся кольцевых окрестностей точек наблюдений снижал мультиколлинеарность предикторов, оцененную по фактору инфляции дисперсии, в среднем 19,16 раза (бутстреп-оценка; 0,95 доверительный интервал: 18,79–19,84) и повышал устойчивость оценок коэффициентов регрессионной модели, измеренную по числу обусловленности, в среднем в 7,52 раза (бутстреп-оценка; 0,95 доверительный интервал: 7,35–7,67) по сравнению с классическим подходом Land Use Buffers к построению пространственных переменных на основе вложенных круговых окрестностей точек наблюдений в серии численных экспериментов (рисунок 3.2).

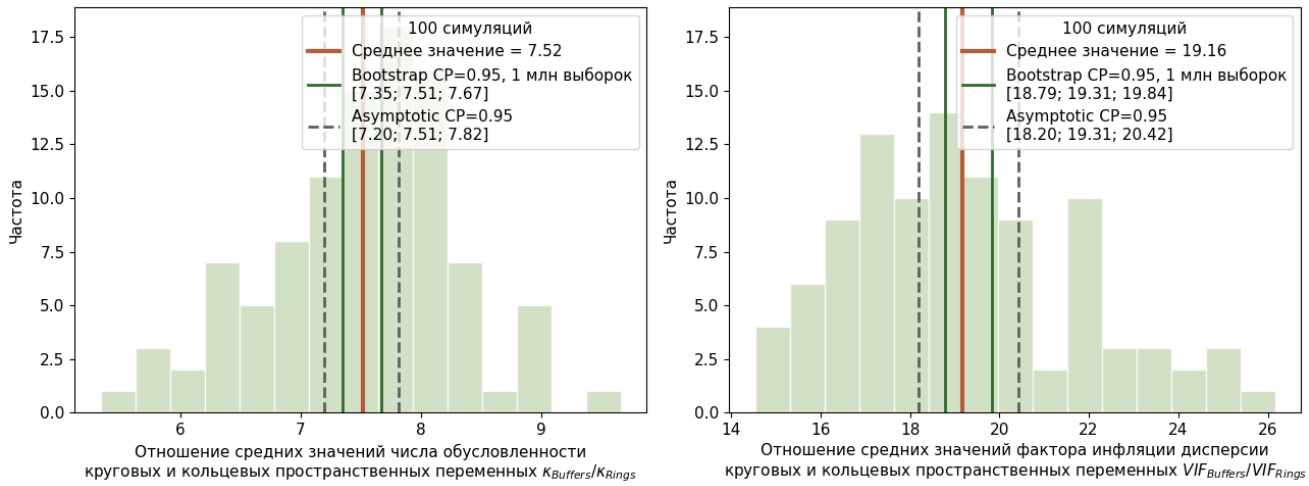


Рисунок 3.2 Распределения отношений средних значений числа обусловленности и фактора инфляции дисперсии круговых и кольцевых пространственных переменных

3.1.3 Результаты численных экспериментов на данных наблюдений

Эксперимент проведен на данных наблюдений потока пыли в снеговом покрове на территории г. Новый Уренгой. В качестве источников примеси на исследуемой территории рассматривалась дорожная сеть. Для каждой точки наблюдения построены предикторы на основе круговых и кольцевых окрестностей, радиусы которых варьировались в диапазоне от 50 до 700 м. Результаты расчета фактора инфляции дисперсии VIF представлены в таблице 3.5.

Таблица 3.5 Расчет фактора инфляции дисперсии VIF

r_k	VIF	(r_{k-1}, r_k)	VIF
50	12,22	(0, 50)	**2,69**
75	26,71	(50, 75)	**3,74**
110	21,29	(75, 110)	**3,03**
160	20,94	(110, 160)	**2,94**
230	19,11	(160, 230)	**2,79**
330	10,35	(230, 330)	**2,22**
700	2,82	(330, 700)	**1,50**

Для предикторов на основе круговых окрестностей значения VIF достигали 26,7 (медианное – порядка 20), при этом большинство признаков имели VIF > 10, что свидетельствовало о сильной линейной зависимости между предикторами. Для

кольцевых предикторов значения VIF существенно ниже: максимальное значение составляет 3,74, медианное – около 2,9, и ни один из признаков не превышает пороговое значение VIF = 10, что указывает на снижение мультиколлинеарности при переходе к кольцевым пространственным предикторам.

Число обусловленности для круговых пространственных переменных составило 16,3, что указывает на умеренную обусловленность матрицы предикторов. В случае кольцевых переменных значение числа обусловленности составило 4,9, что свидетельствует о более устойчивой структуре пространства предикторов на основе кольцевых окрестностей точек наблюдений.

Устойчивость моделей дополнительно была проанализирована для различных конфигураций радиусов. Рассматривались четыре конфигурации:

- плотная линейная {50; 100; 150; …; 700},
- линейная базовая {50; 100; 150; 200; 300; 500; 700},
- логарифмическая базовая {50; 90; 130; 180; 250; 400; 700},
- логарифмическая физически мотивированная {50; 75; 110; 160; 230; 330; 700}.

Для каждой конфигурации оценивались число обусловленности, медианное значение фактора инфляции дисперсии VIF, максимальные значения VIF, а также доля признаков с VIF > 10 (таблица 3.6).

Таблица 3.6 Анализ устойчивости моделей для различных конфигураций радиусов пространственных переменных

Конфигурация радиусов	Тип предикторов	Число предикторов	κ	VIF		Доля VIF > 10
				median	max	
Плотная линейная	Buffer	14	95,50	137,92	375,20	0,93
	Ring	14	**7,80**	**3,48**	**428**	**0,00**
Линейная базовая	Buffer	7	16,92	14,11	25,92	0,71
	Ring	7	**4,71**	**2,80**	**3,10**	**0,00**
Логарифмическая базовая	Buffer	7	14,95	15,00	21,77	0,57
	Ring	7	**4,35**	**2,43**	**3,08**	**0,00**
	Buffer	7	16,27	19,11	26,71	0,86

Логарифмическая физически мотивированная	Ring	7	**4,92**	**2,79**	**3,74**	**0,00**

Для пространственных переменных, построенных на основе круговых окрестностей, для всех рассмотренных конфигураций радиусов наблюдалась высокая мультиколлинеарность предикторов. Значения числа обусловленности находились в диапазоне 15–96, при этом медианные значения VIF превышали 14, а в некоторых конфигурациях превышали 100. Доля предикторов с VIF > 10 превысила 0,9 для плотной линейной конфигурации.

В противоположность этому, предикторы, построенные на основе кольцевых окрестностей точек наблюдений, демонстрировали существенно более устойчивое поведение и стабильно низкую мультиколлинеарность. Во всех конфигурациях радиусов значения числа обусловленности не превышали 8, медианные значения VIF находились в диапазоне 2,4–3,5, а предикторы с VIF > 10 отсутствовали.

Сравнение различных конфигураций радиусов показало, что увеличение плотности радиусов приводит к существенному росту мультиколлинеарности, при этом логарифмическое распределение радиусов не дает значимого выигрыша по сравнению с линейным при одинаковом числе предикторов.

Полученные результаты согласуются с результатами численных экспериментов на синтетических данных. Показано, что структура предикторов оказывает определяющее влияние на устойчивость модели, причем переход от круговых окрестностей к кольцевым является более значимым фактором, чем выбор конкретной конфигурации радиусов. Переход к кольцевым предикторам приводит к существенному снижению мультиколлинеарности и повышению численной устойчивости модели, однако полностью не устраняет линейные зависимости между предикторами.

3.2 Оценка качества моделей

В разделе приводится оценка качества моделей, использующих пространственные предикторы, сформированные на основе круговых (Land Use Buffers) и кольцевых (Land Use Rings) окрестностей точек наблюдений, и модели Land Use Segmentation.

3.2.1 Постановка численного эксперимента

Численный эксперимент проводился на синтетической территории размером 5000×5000 м². На каждой симуляции (N_{sim} = 10) при фиксированных псевдослучайных последовательностях выполнялись следующие шаги:

- генерировались источники примеси;

- рассчитывалось наблюдаемое пространственное распределение примеси по сетке территории с использованием заданной функции вклада 2.8;

- для имитации измерений на территории выбирались точки наблюдений s_i, $i = 1, \ldots, n = 100$. Из вычисленного наблюдаемого пространственного распределения извлекались значения содержания примеси в точках наблюдений $C_i = C(s_i)$ и зашумлялись мультипликативным гауссовским шумом с нулевым средним и стандартным отклонением σ: $C_i^{obs} = C_i(1 + \varepsilon),\ \varepsilon \sim \mathcal{N}(0, \sigma^2)$;

- для моделей Land Use Buffers и Land Use Rings формировались пространственные предикторы на основе характеристик источников примеси внутри круговых и кольцевых окрестностей точек наблюдений соответственно. Для снижения мультиколлинеарности применялась пошаговая фильтрация предикторов по VIF. Переменные с VIF > 5 последовательно исключались из матрицы предикторов;

- модели, описанные в разделах 2.6–2.7, обучались с использованием зашумленных наблюдений C_i^{obs}. Для всех моделей, за исключением модели Land Use Segmentation и стандартной линейной регрессии, выполнялся подбор гиперпараметров с использованием вложенной процедуры GridSearchCV в рамках схемы пространственной кросс-валидации. Для этого точки наблюдений разбивались на пять пространственных групп с использованием алгоритма

k-средних) после чего последовательно проводилось обучение на четырех группах и тестирование на оставшейся группе;

- после подбора гиперпараметров вычислялись прогнозы для всех точек наблюдений относительно наблюдаемого пространственного распределения примеси C_i. Метрики качества моделей вычислялись на основе сравнения предсказанного набора данных со значениями примеси C_i;

- на заключительном этапе выполнялось построение карт пространственного распределения примеси по всей территории.

3.2.2 Результаты численных экспериментов на синтетических данных

Для исследования устойчивости моделей к ошибкам измерений численный эксперимент был проведен при различных значениях стандартного отклонения σ мультипликативного гауссовского шума: $\sigma \in \{0; 0{,}05; 0{,}10; 0{,}20; 0{,}30; 0{,}50\}$. Построена зависимость ошибки моделей *RMSE* от σ (рисунок 3.3). Точность прогноза монотонно снижалась с увеличением σ для всех оцениваемых моделей, при этом скорость снижения точности существенно различалась в зависимости от класса моделей. Регуляризованные линейные модели, в частности гребневая регрессия, продемонстрировали наибольшую устойчивость к зашумленным наблюдениям, поддерживая относительно низкие значения *RMSE* даже при $\sigma = 0{,}5$. Ансамблевые методы показали умеренную устойчивость. Random Forest продемонстрировал сравнительно стабильное поведение при всех σ, в то время как Gradient Boosting показал прогрессивно возрастающую ошибку при более высоких уровнях шума, вероятно, из-за чувствительности к зашумленной подгонке остатков. MLP был наиболее чувствителен к шуму. *RMSE* резко возрастал с σ, сопровождаясь существенно более широкими межквартильными интервалами в ходе моделирования, особенно для модели Land Use Buffers. Такое поведение согласуется с известной нестабильностью нейросетевых моделей в условиях малых наборов данных и наличия шума в измерениях. В отличие от моделей Land Use Buffers и Rings, Land Use Segmentation продемонстрировала наибольшую устойчивость к шуму. Ошибка прогнозирования *RMSE* для Land Use Segmentation

оставалась неизменно ниже, чем у всех моделей Land Use Buffers и Rings на всем диапазоне σ.

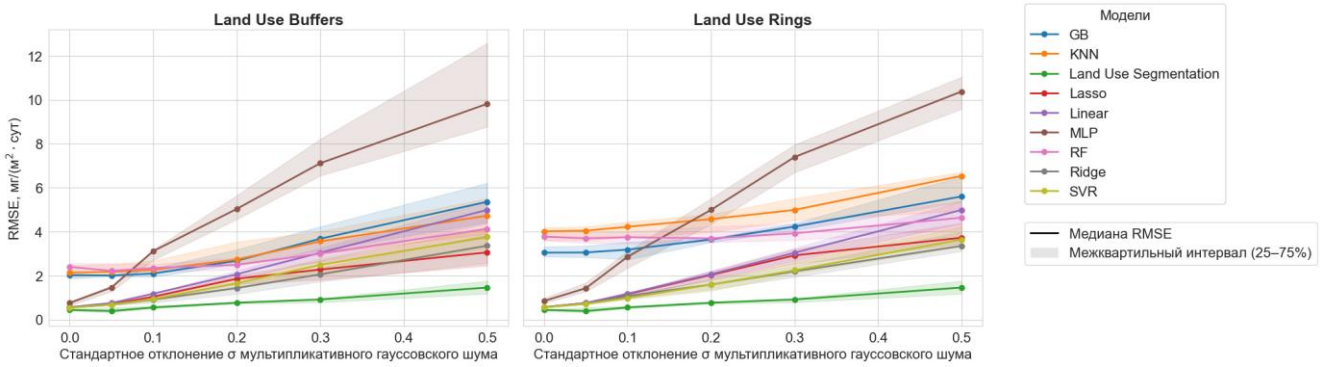


Рисунок 3.3 Зависимость ошибки моделей *RMSE* от стандартного отклонения σ мультипликативного гауссовского шума

Далее было проведено более детальное сравнение моделей на синтетических данных при σ = 0,2. Основные метрики качества моделей приведены в таблице 3.7, дополнительные метрики – в приложении В. Среди Land Use Buffers и Land Use Rings моделей наиболее высокую точность показала модель Ridge. Модель Land Use Segmentation продемонстрировала наибольшую корреляцию между предсказанным и наблюдаемым распределением (*Corr* = 0,996 против 0,993 и 0,991 для моделей Land Use Buffers и Land Use Rings с аппроксиматором Ridge, соответственно) и снижение ошибки прогнозирования (*RMSE* = 1,073 против 1,379 и 1,582 для моделей Land Use Buffers и Land Use Rings с аппроксиматором Ridge, соответственно). Это указывает на более точное воспроизведение пространственной структуры распределения примеси моделью Land Use Segmentation по сравнению с Land Use Buffers и Land Use Rings.

Таблица 3.7 Оценка качества моделей на синтетических данных

Модель Land Use		R^2	*Corr*	*p-value* (*Corr*)	*MAE*	*RMSE*	$Diff_{means}$	*p-value* ($Diff_{means}$)	$Diff_{medians}$	*p-value* ($Diff_{medians}$)
Linear	Buffers	0,967	0,986	0,000	1,542	2,016	0,086	0,434	0,248	0,343
	Rings	0,967	0,986	0,000	1,542	2,016	0,086	0,434	0,248	0,343
Ridge	**Buffers**	**0,984**	**0,993**	**0,000**	**1,066**	**1,379**	0,065	0,450	0,057	0,427
	Rings	**0,979**	**0,991**	**0,000**	**1,231**	**1,582**	0,074	0,441	0,204	0,364
Lasso	Buffers	0,973	0,988	0,000	1,393	1,805	0,053	0,459	0,156	0,394
	Rings	0,967	0,986	0,000	1,522	1,986	0,087	0,432	0,268	0,330
RF	Buffers	0,931	0,969	0,000	1,975	2,822	–0,296	0,274	–0,300	0,262

	Rings	0,860	0,936	0,000	2,945	4,108	−0,514	0,136	–0,268	0,231
GB	Buffers	0,928	0,967	0,000	2,122	2,951	–0,133	0,396	–0,328	0,276
	Rings	0,881	0,942	0,000	2,780	3,804	−0,242	0,311	–0,419	0,243
kNN	Buffers	0,917	0,965	0,000	2,132	3,127	−0,763	0,056	−0,831	0,048
	Rings	0,825	0,933	0,000	3,364	4,615	−1,727	0,000	−1,452	0,002
SVR	Buffers	0,975	0,990	0,000	1,248	1,654	**−0,006**	**0,496**	**0,035**	**0,486**
	Rings	0,980	0,992	0,000	1,184	1,554	**−0,002**	**0,498**	**−0,135**	**0,342**
MLP	Buffers	0,749	0,911	0,000	3,804	5,324	0,088	0,438	0,206	0,331
	Rings	0,787	0,909	0,000	3,825	5,039	0,075	0,443	−0,247	0,350
Land Use Segmentation		**0,990**	**0,996**	**0,000**	**0,831**	**1,073**	**0,055**	**0,456**	**0,066**	**0,439**

После Land Use Buffers и Land Use Rings (Ridge) следовали Rings (SVR) с $R^2 = 0{,}980$; $RMSE = 1{,}554$) и Buffers (SVR) с буферами ($R^2 = 0{,}975$; $RMSE = 1{,}654$). Близкие результаты продемонстрировала также модель Buffers (Lasso) ($R^2 = 0{,}973$; $RMSE = 1{,}805$). Линейная регрессия показала несколько более низкую точность ($R^2 = 0{,}967$; $RMSE = 2{,}016$), однако результаты оставались достаточно стабильными и превосходили метрики ансамблевых методов и нейронной сети.

Для моделей Random Forest и Gradient Boosting наблюдалось снижение качества прогноза. Например, для Random Forest (Rings) R^2 снизился до 0,860, а $RMSE$ увеличилась до 4,108, тогда как для Gradient Boosting (Rings) получены $R^2 = 0{,}881$ и $RMSE = 3{,}804$. Наихудшие результаты были получены для MLP, где значения R^2 составили 0,749–0,787, а ошибка $RMSE$ достигала 5,324. Это может свидетельствовать о том, что малом размере обучающего подмножества нейронная сеть не смогла устойчиво восстановить зависимость между пространственными предикторами и содержанием примеси.

Для $\sigma = 0{,}2$ была построена диаграмма Тейлора с использованием усредненных по симуляциям данных (рисунок 3.4). При $\sigma = 0{,}2$ влияние шума становится достаточно отчетливым. Хотя снижение точности прогноза уже очевидно, качество большинства моделей остается достаточно хорошим, что позволяет оценить различия в поведении моделей. Предложенная модель Land Use Segmentation остается наиболее близким к эталонной точке, что указывает на наилучшее соответствие целевому пространственному распределению при $\sigma = 0{,}2$.

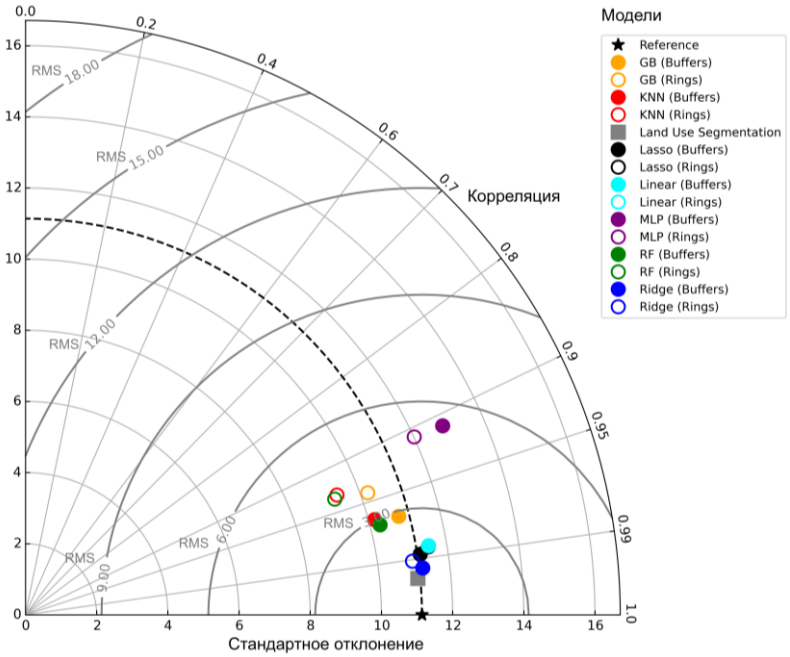


Рисунок 3.4 Диаграмма Тейлора для моделей на синтетических данных

Наконец, выполнялось построение пространственных распределений примеси по сетке территории (рисунок 3.5). Для этого отобранные по минимальной *RMSE* модели Land Use Buffers и Land Use Rings и модель Land Use Segmentation предсказывали содержание примеси в каждом сегменте территории. Карты построены по территории, смоделированной на первой симуляции (sim = 0).

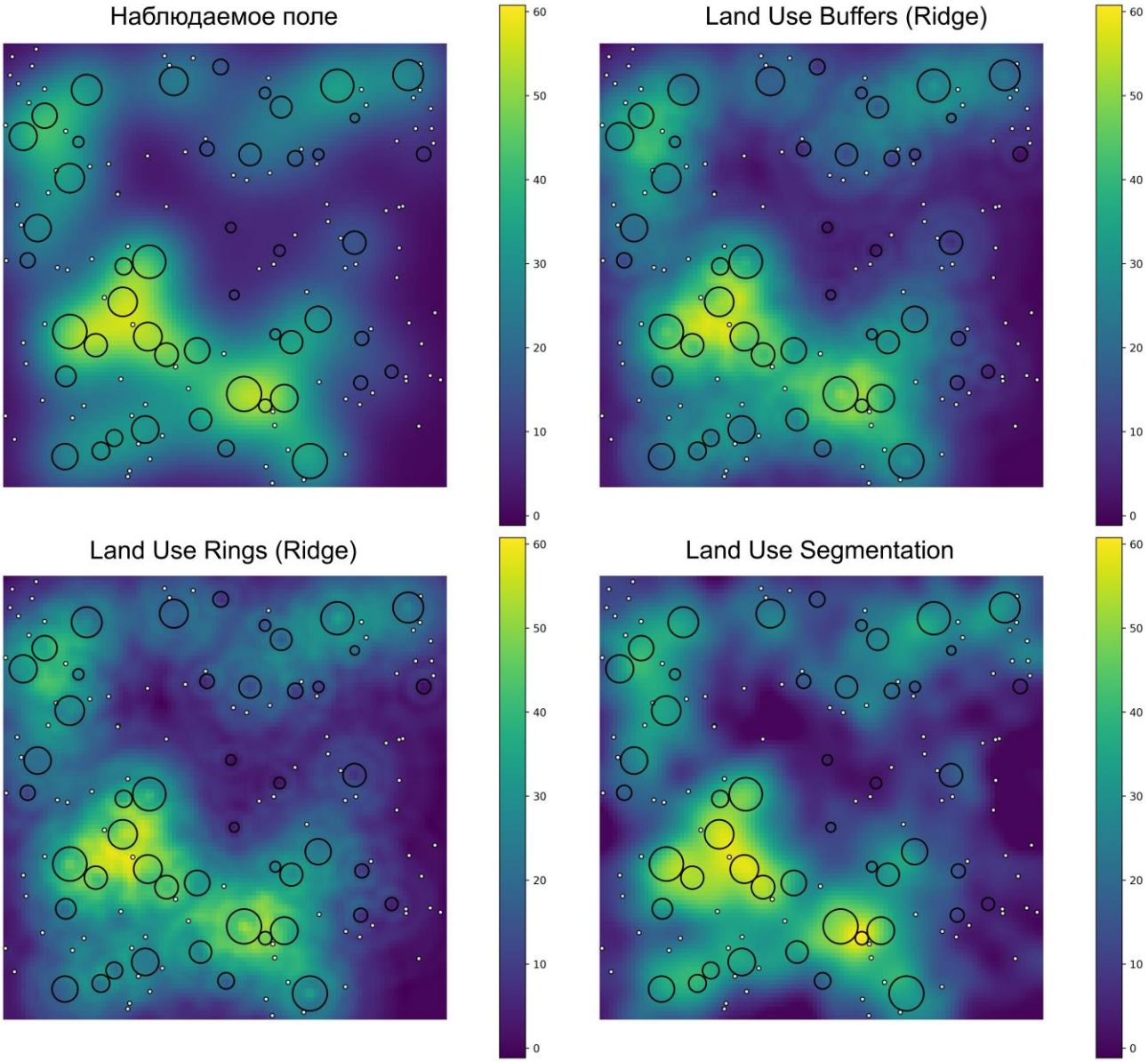


Рисунок 3.5 Пространственное распределение примеси на синтетической территории (иллюстрация для sim = 0 и σ = 0,2)

Пространственное распределение примеси, построенное моделью Land Use Segmentation, имеет меньшую ошибку, чем Land Use Buffers и Land Use Rings, а также характеризуется более гладкой пространственной структурой.

3.2.3 Результаты численных экспериментов на данных наблюдений

В настоящей работе также выполнен сравнительный анализ моделей Land Use Buffers, Land Use Rings и Land Use Segmentation на данных наблюдений потока пыли в снеговом покрове на территории г. Новый Уренгой.

Для каждой точки наблюдения были построены наборы круговых и кольцевых окрестностей с радиусами $r_k \in \{50;\ 75;\ 110;\ 160;\ 230;\ 330;\ 700\}$ м, соответствующих: окружностям $(0,\ r_k)$ и кольцам $(r_{k-1},\ r_k)$. Для каждой круговой/кольцевой окрестности в качестве предиктора вычислялась площадь пересечения круговой/кольцевой окрестности и источников примеси (дорог).

Все предикторы были стандартизированы перед построением моделей Land Use Buffers и Land Use Rings. Было рассмотрено несколько алгоритмов регрессии: гребневая регрессия и Elastic Net, случайный лес, градиентный бустинг, *k*-ближайших соседей. Этот набор включал как линейные, так и нелинейные модели, избегая при этом чрезмерной сложности модели. Цель состояла в том, чтобы построить интерпретируемые и воспроизводимые базовые (baseline) модели, репрезентативные для классического подхода Land Use.

Качество моделей оценивалось в режиме разбиения набора данных на обучающее и тестовое подмножество в соотношении 80%:20% (random train/test split).

В таблице 3.8 приведены результаты оценки качества построенных моделей. Наилучшее качество продемонстрировала модель Land Use Segmentation, для которой коэффициент детерминации составил $R^2 = 0{,}723$, коэффициент корреляции $Corr = 0{,}890$, а значения ошибок $MAE = 0{,}014$ и $RMSE = 0{,}019$ наименьшие среди всех исследованных моделей. Среди моделей Land Use Buffers и Land Use Rings наилучшие результаты были получены для архитектур ElasticNet и Ridge. Для модели ElasticNet значения R^2 составили 0,470 и 0,468 для Buffers и Rings соответственно, а коэффициенты корреляции достигли 0,727 и 0,725. Для Ridge аналогичные показатели составили $R^2 = 0{,}465$ и $R^2 = 0{,}402$ при коэффициентах корреляции 0,704 и 0,658 соответственно.

Нелинейные модели Random Forest, Gradient Boosting, kNN показали более низкие значения коэффициента детерминации ($R^2 = 0{,}028$–$0{,}290$) и более высокие ошибки прогнозирования ($RMSE = 0{,}030$–$0{,}035$). Несмотря на умеренные значения коэффициентов корреляции ($Corr = 0{,}449$–$0{,}589$), данные модели демонстрируют

меньшую способность к воспроизведению общего паттерна пространственного распределения примеси по сравнению с линейными моделями.

Дополнительные метрики качества моделей приведены в приложении В.

Таблица 3.8 Оценка качества моделей на данных наблюдений потока пыли в снеговом покрове г. Новый Уренгой

Модель Land Use		R^2	*Corr*	*p-value* (*Corr*)	*MAE*	*RMSE*	$Diff_{means}$	*p-value* ($Diff_{means}$)	$Diff_{medians}$	*p-value* ($Diff_{medians}$)
Ridge	Buffers	0,465	0,704	0,000	0,021	0,026	0,005	0,249	0,005	0,324
	Rings	0,402	0,658	0,000	0,022	0,027	0,006	0,217	0,008	0,253
ElasticNet	**Buffers**	**0,470**	**0,727**	**0,000**	**0,021**	**0,026**	0,007	0,193	0,007	0,278
	Rings	**0,468**	**0,725**	**0,000**	**0,021**	**0,026**	0,007	0,186	0,007	0,278
RF	**Buffers**	0,223	0,539	0,002	0,024	0,031	**0,003**	**0,381**	**0,000**	**0,500**
	Rings	0,270	0,589	0,001	0,023	0,030	**0,004**	**0,316**	**–0,001**	**0,488**
GB	Buffers	0,129	0,479	0,005	0,026	0,033	0,003	0,342	–0,003	0,355
	Rings	0,028	0,449	0,009	0,026	0,035	0,003	0,362	–0,008	0,268
kNN	Buffers	0,290	0,553	0,001	0,024	0,030	0,002	0,414	0,000	0,497
	Rings	0,109	0,480	0,004	0,025	0,033	0,006	0,227	0,009	0,289
Land Use Segmentation		**0,723**	**0,890**	**0,000**	**0,014**	**0,019**	–0,009	0,145	–0,009	0,237

Для визуализации результатов моделей на данных наблюдений потока пыли была построена диаграмма Тейлора (рисунок 3.6). Среди моделей Land Use Buffers и Land Use Rings наилучшее качество продемонстрировали регуляризованные линейные методы Elastic Net и Ridge. Для них наблюдаются наибольшие значения корреляции (от 0,658 до 0,727) при сравнительно небольшом отклонении от наблюдаемого распределения. Среди нелинейных моделей наиболее согласована с наблюдаемыми данными модель Random Forest, которая, однако, характеризуется несколько большей ошибкой, чем линейные модели. Модели kNN и Gradient Boosting расположены дальше от эталонной точки и имеют более высокие значения *RMS*, что указывает на снижение точности воспроизведения пространственной структуры данных. Различия в точности между Land Use Buffers и Land Use Rings невелики: для большинства моделей наблюдается близкое положение соответствующих маркеров на диаграмме. Модель Land Use Segmentation показала

наибольшую корреляцию и наименьшую ошибку прогноза, при этом ее стандартное отклонение отличается от эталонного, что может свидетельствовать о смещении амплитуды прогноза относительно наблюдаемых значений.

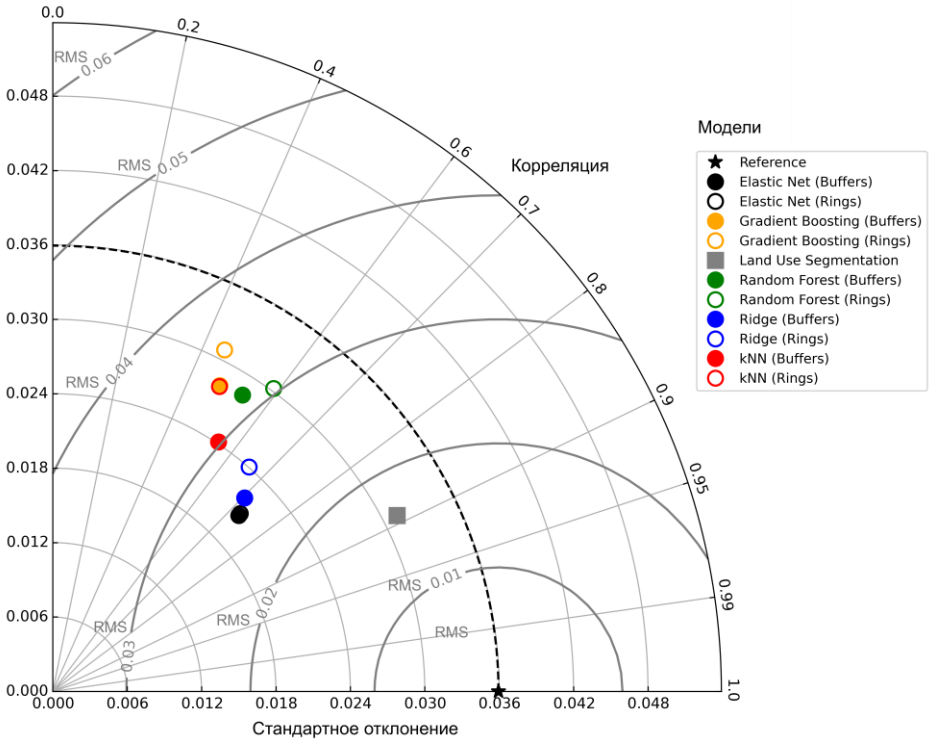
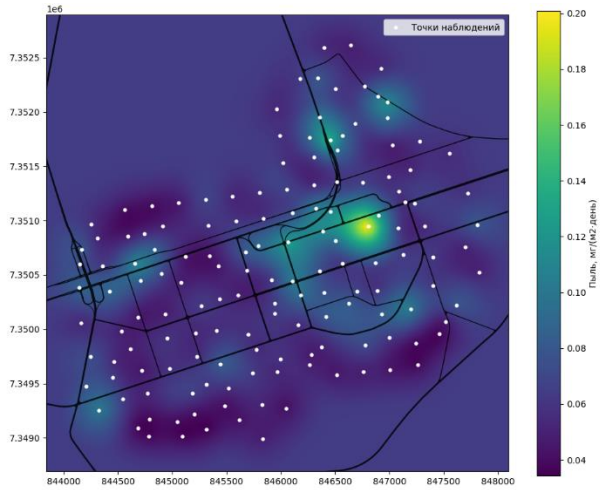
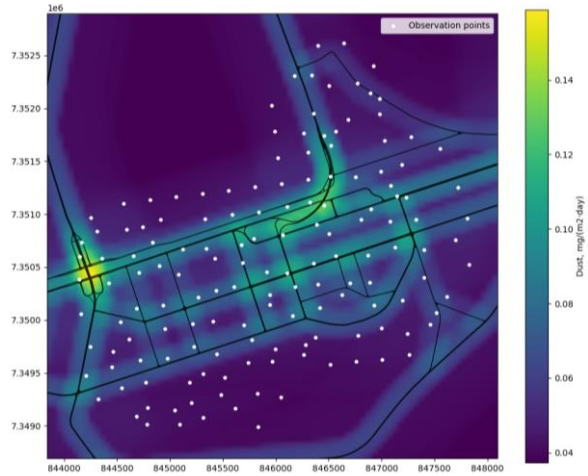


Рисунок 3.6 Диаграмма Тейлора для моделей на данных наблюдений потока пыли в снеговом покрове г. Новый Уренгой

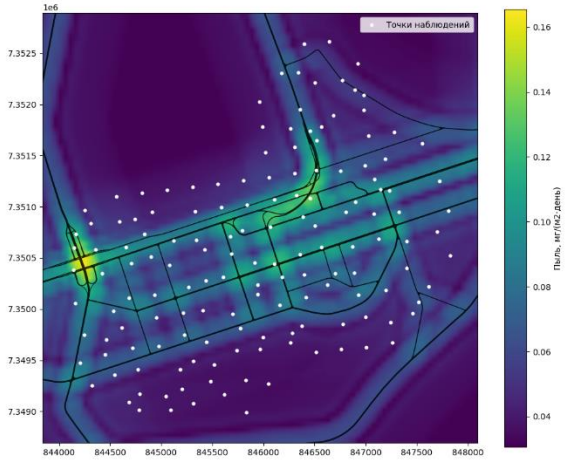
Для построения пространственных распределений использовалась регулярная сетка с шагом 50 м. Обученные на данных наблюдений модели прогнозировали содержание потока пыли в узлах сетки (рисунок 3.7). Для сравнения с моделями Land Use была построена карта пространственного распределения методом обычного кригинга (Ordinary Kriging) (рисунок 3.7, а). Этот подход отражает исключительно пространственную автокорреляцию наблюдений и не учитывает пространственные факторы формирования распределения примеси на территории. Полученное распределение представляет собой интерполяцию измеренных значений между точками наблюдений без явной связи с расположением дорожной сети как потенциальных источников примеси.



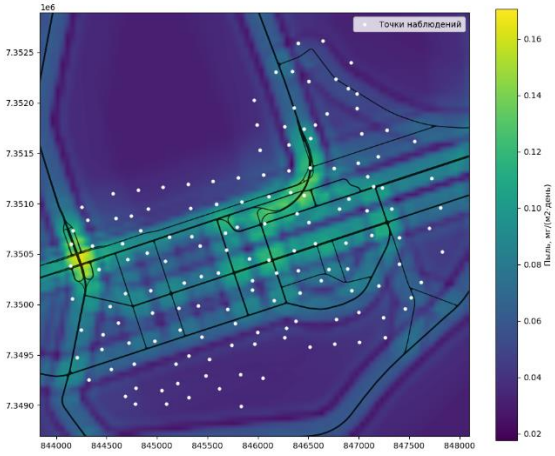
(а) Кригинг



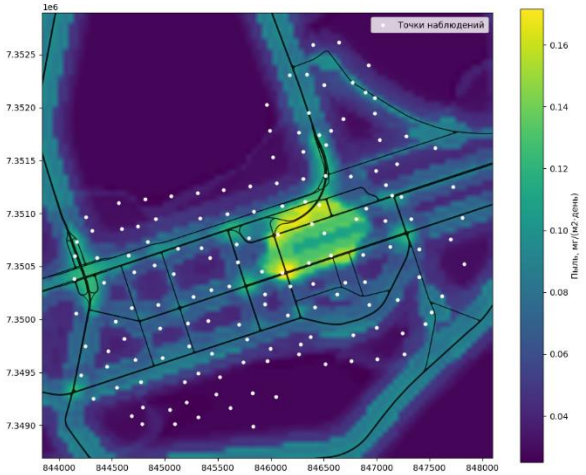
(б) Land Use Segmentation



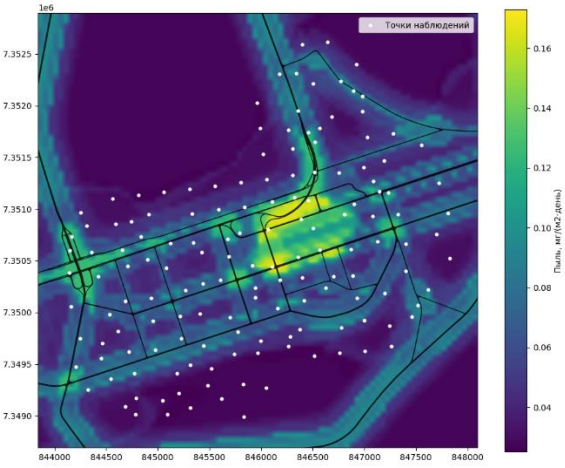
(в) Лучшая линейная модель
Land Use Buffers (ElasticNet)



(г) Лучшая линейная модель
Land Use Rings (ElasticNet)



(д) Лучшая нелинейная модель
Land Use Buffers (Random Forest)



(е) Лучшая нелинейная модель
Land Use Rings (Random Forest)

Рисунок 3.7 Пространственное распределение потока пыли в снеговом покрове г. Новый Уренгой

В отличие от кригинга, модели Land Use учитывают пространственные характеристики источников примеси. В модели Land Use Buffers и Rings информация о дорожной сети включалась через систему пространственных

предикторов, учитывающих площадь дорог в круговых/кольцевых окрестностях узлов сетки, в модель Land Use Segmentation – через функцию вклада.

Полученные карты (рисунок 3.7, б–е) демонстрируют пространственное распределение потока пыли в снеговом покрове, согласованное с конфигурацией транспортной инфраструктуры на территории г. Новый Уренгой. Наиболее высокое качество среди исследованных моделей продемонстрировала Land Use Segmentation (рисунок 3.7, б), показавшая минимальное значение ошибки. Распределение характеризуется плавной пространственной структурой без выраженных случайных локальных колебаний. При этом модель обеспечивает достаточно высокую степень детализации и выделяет основные пространственные закономерности распространения примеси. Модели Land Use Buffers и Land Use Rings также корректно восстанавливают пространственное распределение примеси на территории, однако характеризуются несколько более выраженной локальной вариабельностью (рисунок 3.7, в–е). Полученные карты имеют менее сглаженный характер и содержат больше локальных неоднородностей. Несмотря на это, общая пространственная структура воспроизводится моделями Land Use Buffers и Land Use Rings достаточно хорошо, а увеличение ошибки по сравнению с моделью Land Use Segmentation остается относительно небольшим.

Сравнение линейных и нелинейных моделей показало различия в характере восстанавливаемых пространственных полей. Нелинейные модели (например, Random Forest, рисунок 3.7, д, е) формируют более неоднородную и «шумную» пространственную структуру по сравнению с линейными моделями (Linear Regression, Ridge, Elastic Net, рисунок 3.7, в, г). Вероятно, это связано с более высокой гибкостью нелинейных моделей и их способностью выделять локальные особенности в пространственных данных. При этом нелинейные модели позволяют более явно выделять зоны локального накопления примеси. На построенных картах (рисунок 3.7, д, е) наблюдаются более выраженные области повышенных значений потока пыли в районах с наиболее интенсивной транспортной активностью, включая крупные транспортные узлы города Новый Уренгой и территорию железнодорожного вокзала.

Представленные далее результаты иллюстрируют применение разработанных моделей на различных территориях и для различных депонирующих сред. Иллюстрации демонстрируют применимость предложенных моделей к данным наблюдений и дополняют выводы, полученные в ходе численных экспериментов на синтетических данных.

Применение классического LU подхода в г. Новый Уренгой

В каждой точке наблюдения строился набор круговых окрестностей зон радиусами 100, 200, 300, 400 и 500 м. Было получено 10 пространственных переменных: r100, …, r500 – площади пересечений дорог основного типа (roads) с круговыми окрестностями с радиусами 100, …, 500 м соответственно; sr100, …, sr500 – площади пересечений дорог второстепенного типа (subroads) с круговыми окрестностями с радиусами 100, …, 500 м соответственно. В итоговый набор вошли предикторы r100, r500 и sr500, для которых значения VIF не превышали 3. Для выбранных предикторов также учитывались их квадратичные члены.

Тестировались четыре модели: стандартная LUR модель; LUR модель с добавлением географических координат точек наблюдений (LUR+XY); модель на основе MLP; MLP модель с добавлением координат точек наблюдений (MLP+XY). Оценка качества моделей выполнялась на тестовом подмножестве с использованием стандартных метрик качества прогноза: коэффициента корреляции, *MAE*, *RMSE* и *RRMSE*. В таблице 3.9 представлены результаты оценки качества моделей.

Таблица 3.9 Оценка качества моделей

Модель	*Corr*	*MAE*	*RRMSE*	*RMSE*	*SD**
LUR	0,66	10,39	1,21	16,61	3,58
LUR+XY	0,69	10,13	1,09	16,3	3,96
MLP	0,78	8,97	0,85	13,82	6,58
MLP+XY	0,81	8,64	0,7	12,82	7,67

Примечание: *SD** – стандартное отклонение исходных данных = 16,52.

Наилучшее качество (максимальный коэффициент корреляции и минимальные значения ошибок прогноза) продемонстрировала модель MLP с

исходными шестью предикторами, к которым добавили координаты мест отбора проб. Стандартная LUR модель показала самую низкую продуктивность, в то время как добавление координат точек наблюдений позволило незначительно улучшить качество прогноза: коэффициент корреляции модели LUR+XY 0,69 против 0,66 у исходной LUR модели. Тем не менее, качество модели LUR+XY ниже, чем качество моделей на основе MLP. Добавление координат мест отбора проб к предикторам MLP также позволило дополнить модель геостатистической информацией и улучшить ее качество: коэффициент корреляции равен 0,78 для MLP против 0,81 для MLP+XY.

Для восстановления полей поверхностного распределения пыли на исследуемой территории была построена регулярная сетка с шагом 50 м (рисунок 3.8). В качестве базового метода использовался классический геостатистический подход – кригинг, а также четыре модели на основе LUR и MLP. Полученные поля примесей показали, что геостатистический метод позволяет выявить лишь общие пространственные тренды и не дает представления о связи между содержанием примеси и источниками этой примеси. В противоположность кригингу модели с предикторами на основе LU подхода продемонстрировали более точное и интерпретируемое описание распределения примеси с учетом расположения источников загрязнения. При этом поля поверхностного распределения пыли, восстановленные моделями на основе MLP, имеют более высокую детализацию по сравнению с линейными моделями. Результаты опубликованы в [4].

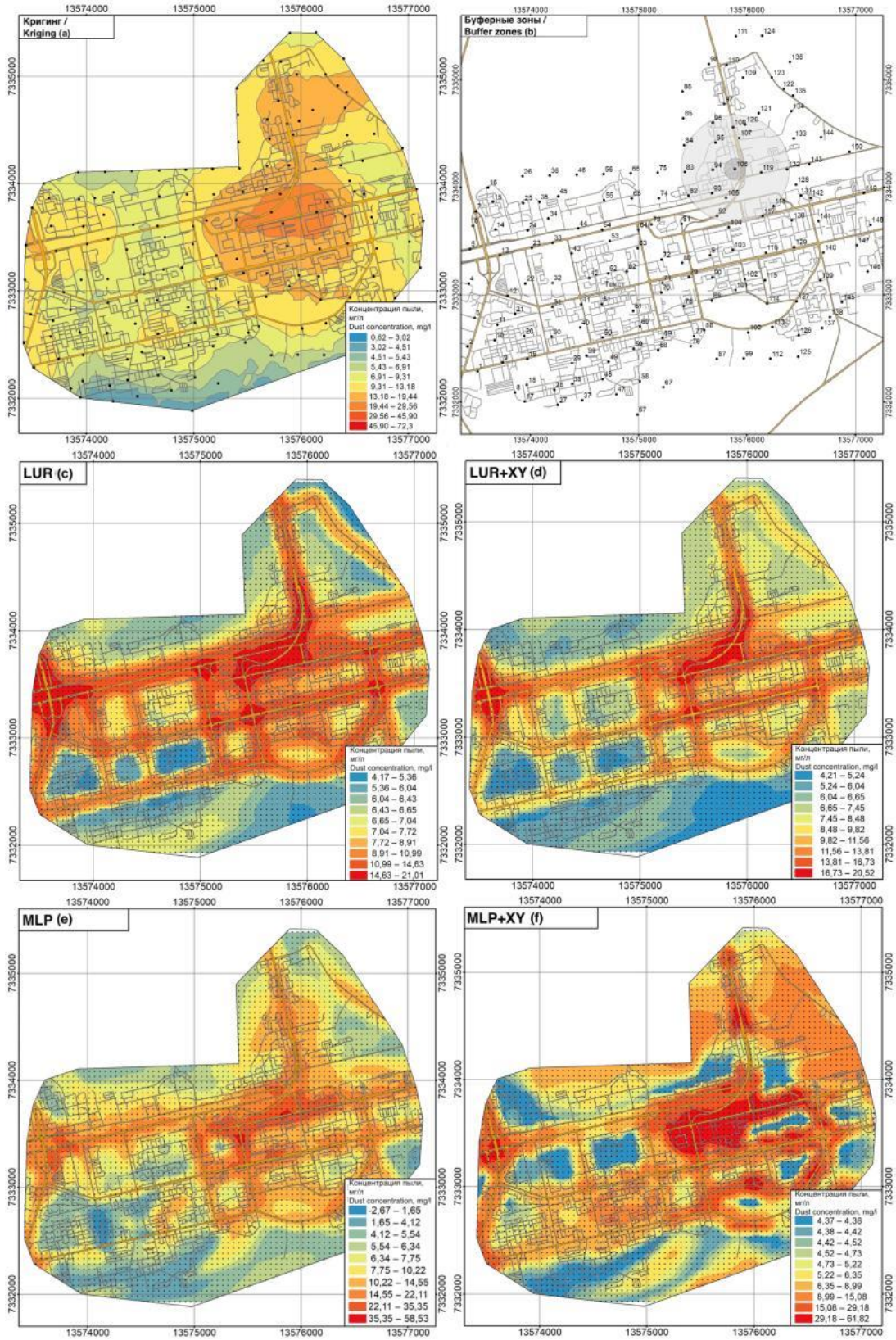


Рисунок 3.8 Моделирование поверхностного распределения пыли в снеговом покрове: (a) – результат кригинга, (b) – пример построения круговых окрестностей радиусами 100 и 500 м для точки наблюдения № 106, (c)–(f) – результаты моделей на основе LUR и MLP

Применение классического LU подхода в г. Тарко-Сале

Для каждой точки наблюдения вычислялось шесть пространственных переменных: r100, r200 и r500 – площади дорог; sr100, sr200 и sr500 – площади парковок, попавших в круговые окрестности с радиусами 100, 200 и 500 м, соответственно. После анализа парных коэффициентов корреляции полученных переменных в модель LUR вошли три предиктора: r100, r500, sr500.

Моделирование выполнялось с использованием двух подходов: стандартной модели LUR и гибридной модели LUR с регрессионным кригингом остатков (LUR+RK). В гибридной модели к прогнозу LUR добавлялась пространственно сглаженная оценка остатков, полученная методом кригинга с использованием экспоненциальной вариограммы с учетом анизотропии.

Результаты оценки точности моделей приведены в таблице 3.10. Для обоих элементов применение регрессионного кригинга позволило повысить точность прогноза по всем используемым метрикам. Так, коэффициент корреляции между измеренными и предсказанными значениями увеличился на 17% для марганца и на 7% для никеля, при одновременном снижении относительной среднеквадратической ошибки на 10% и для никеля, и для марганца.

Таблица 3.10 Оценка качества моделей

Модель	*Corr*	*MAE*	*RRMSE*	*RMSE*	*NRMSE*
Mn_LUR	0,444	74,228	0,660	118,085	1,052
Mn_LUR_Kriging	0,616	64,547	0,603	112,467	1,002
Ni_LUR	0,524	9,311	0,470	11,770	1,152
Ni_LUR_Kriging	0,592	8,560	0,418	11,474	1,123

Диаграмма Тейлора (рисунок 3.9) демонстрирует, что наилучшей предсказательной способностью обладает модель LUR с RK для никеля. Обе модели LUR с RK обладают большим коэффициентом корреляции и меньшим среднеквадратичным значением по сравнению с моделями LUR.

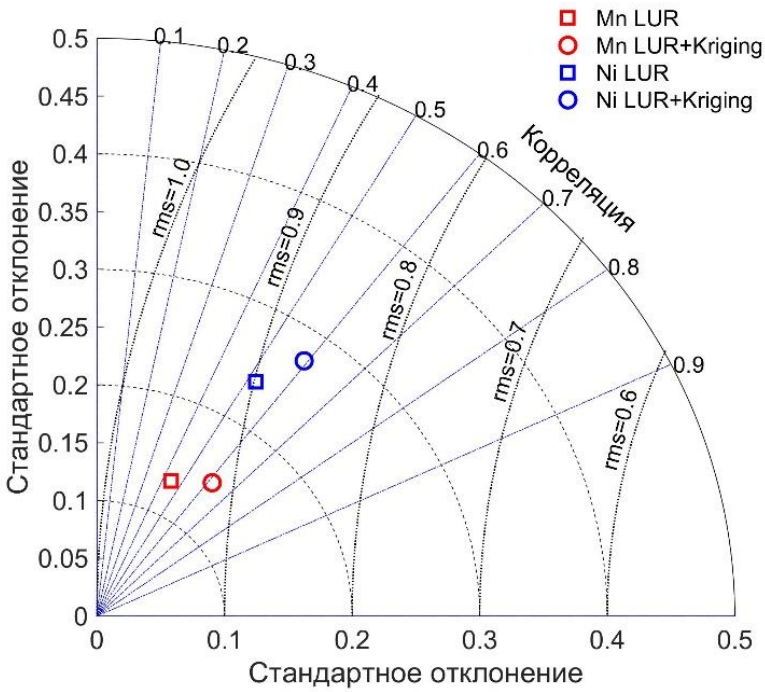


Рисунок 3.9 Диаграмма Тейлора

Полученные поля пространственного распределения марганца и никеля в верхнем слое почвы г. Тарко-Сале показали, что модели LUR и LUR+RK позволяют воспроизводить лишь часть пространственных особенностей распределения марганца и никеля, выявляемых геостатистическим методом (рисунок 3.10). Добавление кригинга остатков повышает согласованность модельных и измеренных данных, однако в целом вклад антропогенных источников, описываемых LU переменными, оказался статистически незначимым для рассматриваемых элементов. Результаты опубликованы в [14].

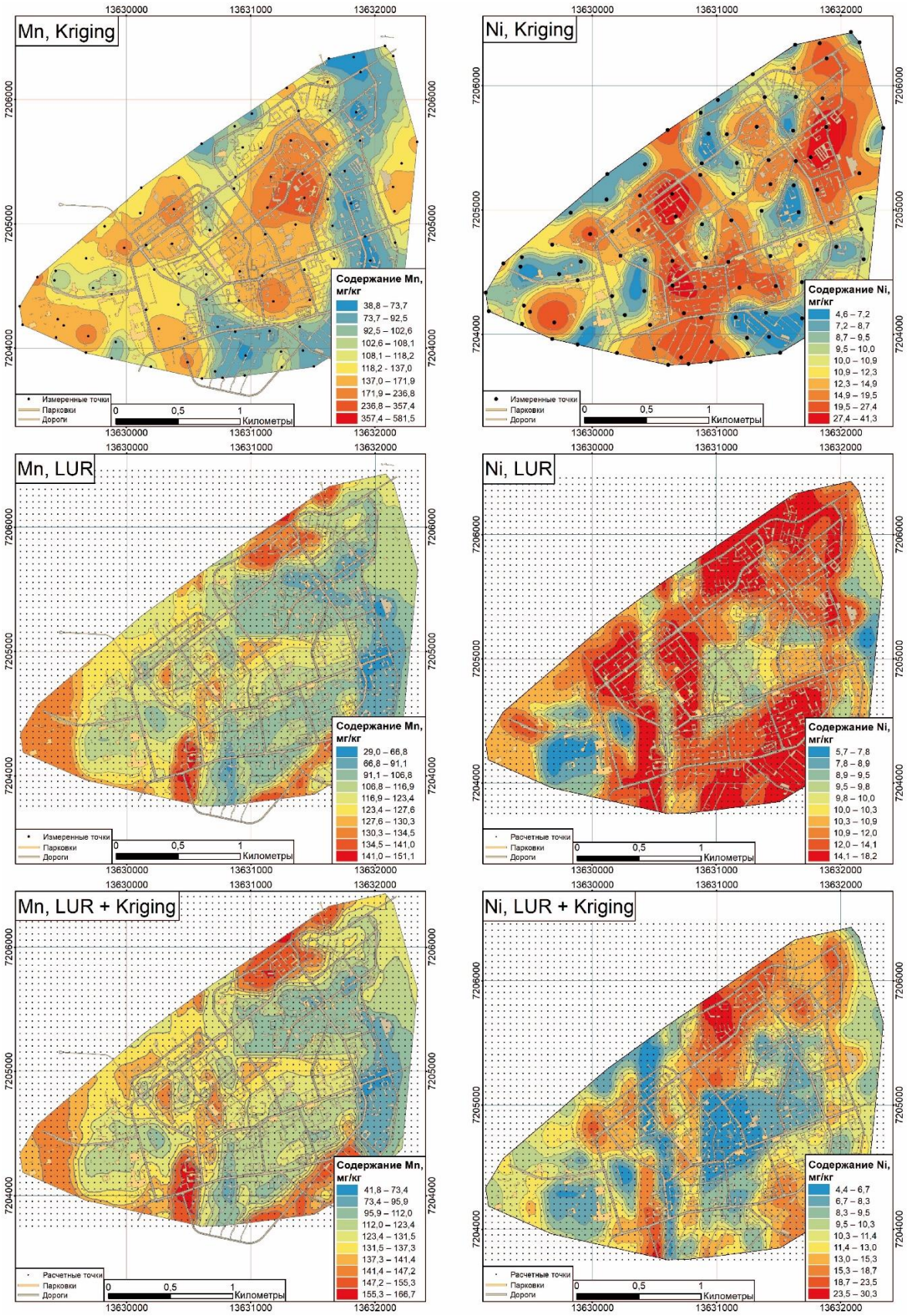


Рисунок 3.10 Результаты моделирования содержания марганца и никеля в верхнем слое почвы г. Тарко-Сале

Применение подхода LU Rings в г. Новый Уренгой

Пространственные переменные формировались на основе непересекающихся кольцевых окрестностей, заданных парами радиусов (0, 100), (100, 200), (200, 300), (300, 400) и (400, 500) м с центрами в точках наблюдений. В результате был получен набор из 10 предикторов, по 5 для каждого типа дорог (основных и второстепенных). По результатам оценок VIF для моделирования были оставлены четыре предиктора: *r*(0, 100), *r*(200-300), *r*(400-500), *sr*(0-100).

Были протестированы шесть моделей: классическая LUR, MLP, и сверточная нейронная сеть (CNN) с использованием отобранных предикторов и те же модели LUR, MLP и CNN с добавлением к набору предикторов координат точек наблюдений. Качество моделей оценивалось по совокупности метрик, включая *MAE*, *RMSE*, *RRMSE*, *NRMSE*, коэффициент корреляции и индексы согласия Уиллмотта (таблица 3.11). Модели на основе ИНС превзошли классическую регрессионную модель по всем основным показателям точности. Наилучшие результаты продемонстрировали модели MLP и CNN с добавлением географических координат, для которых средняя абсолютная ошибка оказалась ниже на 26%, а *RMSE* – ниже на 19% по сравнению с лучшей регрессионной моделью. При этом индексы *IA* и *IA*2 были выше на 3% и 14% соответственно, а коэффициент корреляции был выше на 4%.

Таблица 3.11 Метрики качества моделей

Модель	*MAE*	*RRMSE*	*RMSE*	*NRMSE*	*IA*	*IA*2	*Corr*	*SD*
LUR	2,8	1,7	3,2	0,67	0,88	0,62	0,82	4,9
LUR+XY	2,7	0,8	3,2	0,67	0,89	0,65	0,82	5,1
MLP	2,2	0,7	2,7	0,56	0,92	0,71	0,86	4,7
MLP+XY	2,2	0,7	2,6	0,54	0,92	0,71	0,85	4,6
CNN	2,4	1,4	2,8	0,59	0,90	0,68	0,84	4,3
CNN+XY	2,0	0,7	2,7	0,56	0,92	0,74	0,86	4,7

Визуализация с использованием диаграммы Тейлора (рисунок 3.11) подтвердила близость нейросетевых моделей к наблюдаемым данным по всем трем характеристикам: стандартному отклонению, коэффициенту корреляции и

среднеквадратичной ошибке. Все модели показали сопоставимые значения стандартного отклонения, при этом модели MLP и CNN+XY обладали наименьшими ошибками и наибольшей корреляцией с измеренными значениями.

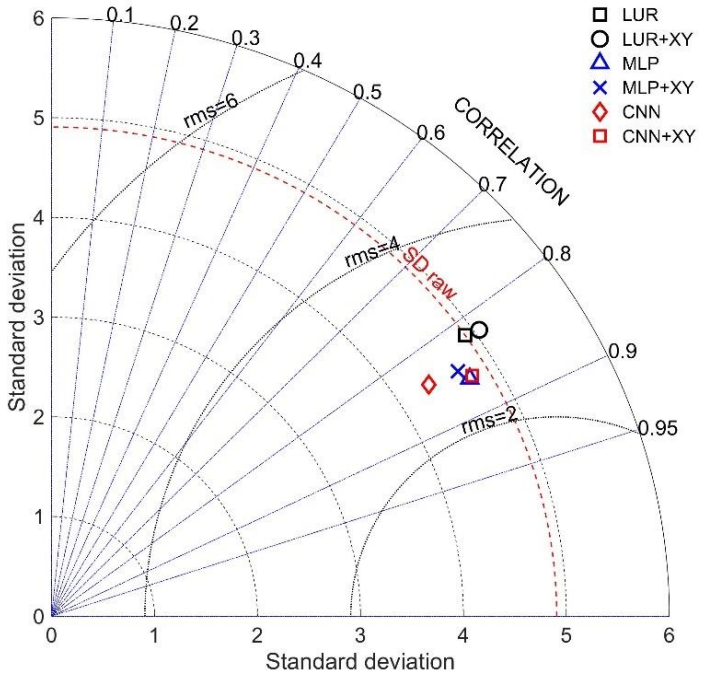


Рисунок 3.11 Диаграмма Тейлора

Наконец, была построена регулярная сетка с шагом 50 м, по которой выполнено восстановление пространственного распределения концентрации пыли (рисунок 3.12). Полученные карты показали выраженную зависимость содержания пыли от расположения дорожной сети и продемонстрировали возможность количественной оценки зоны влияния автомобильных дорог на загрязнение снегового покрова. Результаты опубликованы в [1].

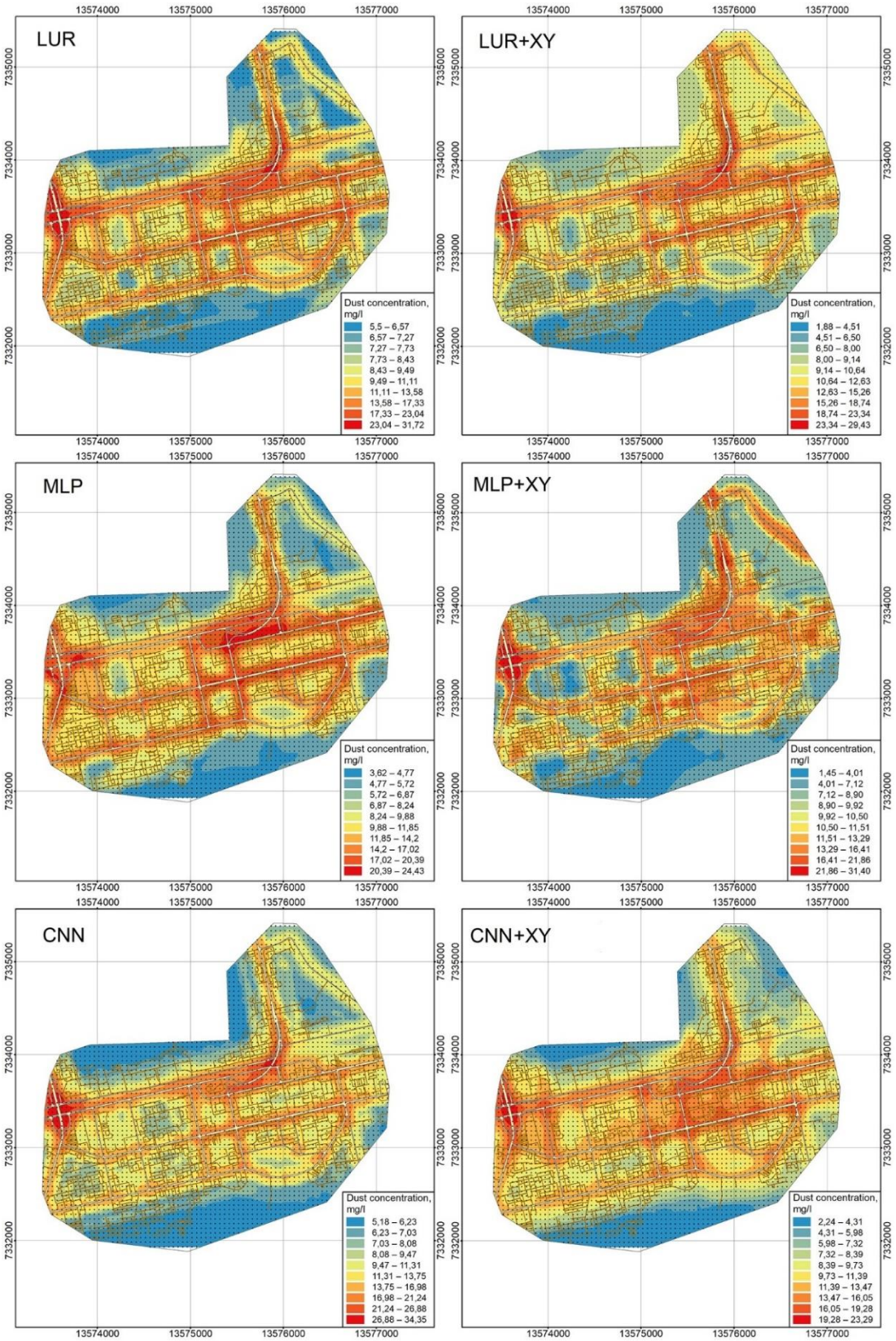


Рисунок 3.12 Карты распределения содержания пыли в снеговом покрове для моделей

Сравнение классического LU подхода и подхода LU Rings в г. Мурманск

Наконец, выполнено контролируемое сравнение классического LU подхода и подхода LU Rings к построению пространственных переменных на одном и том же наборе данных о содержании марганца в поверхностных городских отложениях г. Мурманск. Набор данных включал 40 точек наблюдений, из которых 30 использовались для обучения модели и 10 – для независимого тестирования.

Для каждой точки наблюдения были построены круговые окрестности с радиусами 100, 200, 300, 400 и 500 м и соответствующие им кольцевые окрестности с радиусами (0, 100), (100, 200), …, (400, 500) м. Вычислялись площади пересечений с магистральными и второстепенными дорогами. Таким образом, для каждого подхода было получено по 10 пространственных переменных.

Сравнивались две модели: LUR и MLP с одинаковой архитектурой для обоих наборов предикторов (таблица 3.12). Модели с предикторами на основе кольцевых окрестностей точек наблюдений продемонстрировали более высокую корреляцию между прогнозируемыми и измеренными значениями (0,78 против 0,72 для MLP и 0,65 против 0,64 для LUR). Значения ошибок для линейной и нейросетевой моделей при переходе к кольцевым окрестностям точек наблюдений менялись незначительно.

Таблица 3.12 Оценка качества моделей

Модель	*MAE*	*RMSRE*	*RMSE*	*IA*	*Corr*	*SD*
Стандартный LUR	128,3	0,31	148,6	0,48	0,64	31,7
LUR Rings	130,0	0,31	148,8	0,48	0,65	29,9
Стандартный LU+MLP	121,5	0,26	142,1	0,58	0,72	42,8
MLP Rings	120,7	0,29	141,6	0,54	0,78	35,5

Для каждой модели с использованием тех же предикторов методом кригинга были восстановлены значения в точках регулярной сетки с шагом 100 м. Построенные поля примесей приведены на рисунке 3.13.

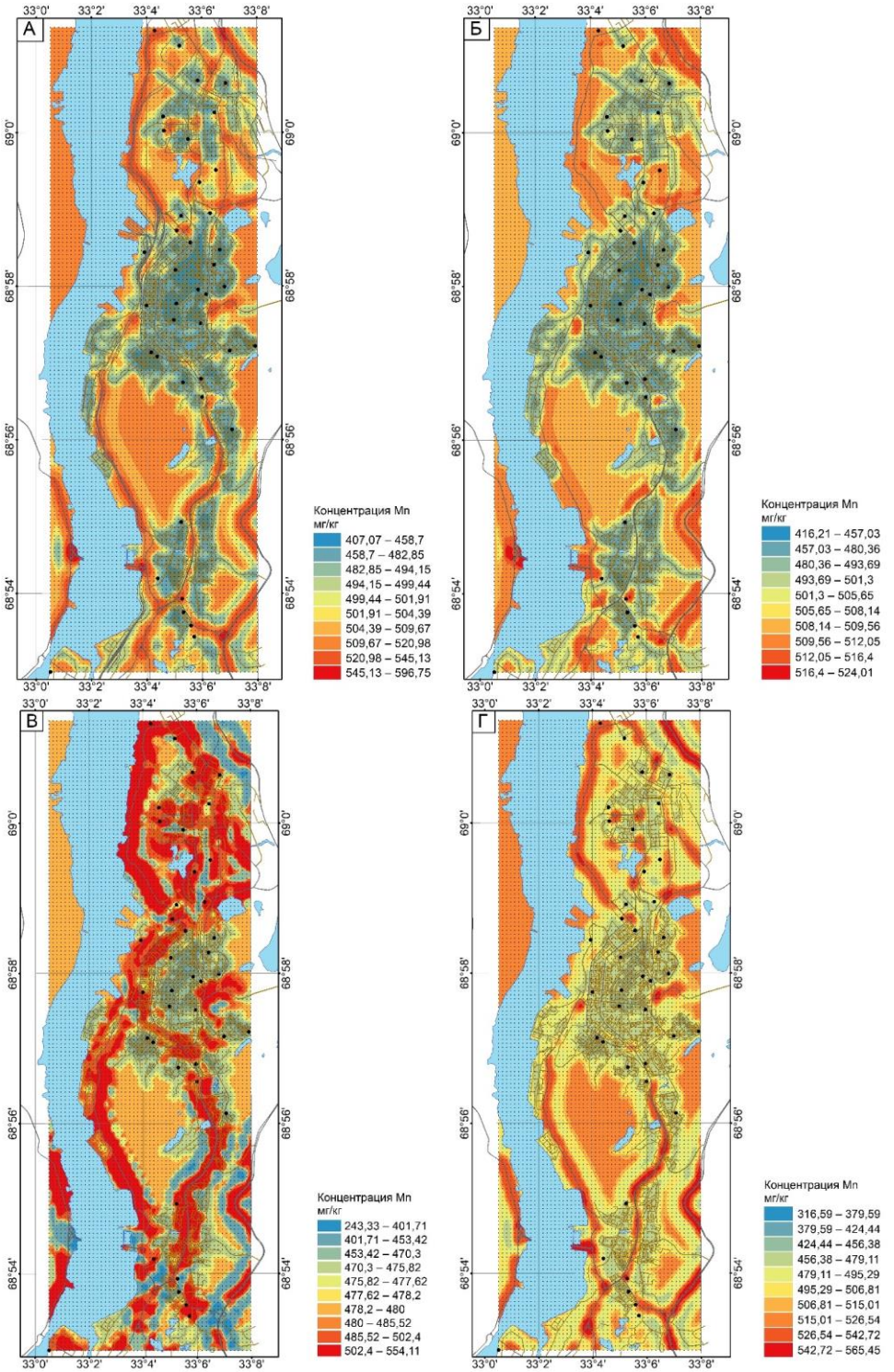


Рисунок 3.13 Распределение марганца в г. Мурманск (А – стандартный LUR, Б – LUR Rings, В – Стандартный LU+MLP, Г– MLP Rings

Таким образом, использование кольцевых пространственных переменных обеспечило более низкую мультиколлинеарность предикторов и более высокую корреляцию прогнозируемых значений с наблюдаемыми, что указывает на более устойчивую пространственную структуру модели. Результаты опубликованы в [9].

3.3 Применение пермутационного подхода к оценке моделей

3.3.1 Постановка численного эксперимента

Для анализа различий между асимптотическим и пермутационным подходами был проведен численный эксперимент на синтетических данных. Рассматривалась линейная модель вида

$$Y = \rho X + \sqrt{1-\rho^2}\varepsilon,$$

где ρ – значение коэффициента корреляции в генеральной совокупности, $X \sim \mathcal{N}(0,1)$ и $\varepsilon \sim \mathcal{N}(0,1)$ – независимые нормально распределенные случайные величины с нулевым средним и единичной дисперсией.

Эксперимент повторялся для различных размеров набора данных $n \in \{8; 10; 12; 15; 20; 25; 30; 40; 60; 100\}$ и силы связи $\rho \in \{0; 0{,}2; 0{,}35; 0{,}5\}$. Для каждой комбинации значений n и ρ строилось 500 независимых реализаций модели. Уже при $n > 30$ полный перебор 2^n конфигураций становится вычислительно недостижимым из-за роста числа комбинаций. По этой причине в настоящей работе генерировалось 100000 случайных перестановок, что позволяло получить устойчивые оценки *p-value* при разумных затратах вычислительных ресурсов.

Для каждой реализации вычислялись асимптотический *p-value* (*p_asymp*) и пермутационный *p-value* (*p_perm*) для трех выбранных статистик. Для разности средних и медиан пермутационный *p-value* сравнивался с асимптотическим *p-value* теста Колмогорова-Смирнова.

3.3.2 Результаты численных экспериментов на синтетических данных

На рисунке 3.14 представлена зависимость среднего абсолютного расхождения между асимптотическими и перестановочными *p-value* от размера набора данных.

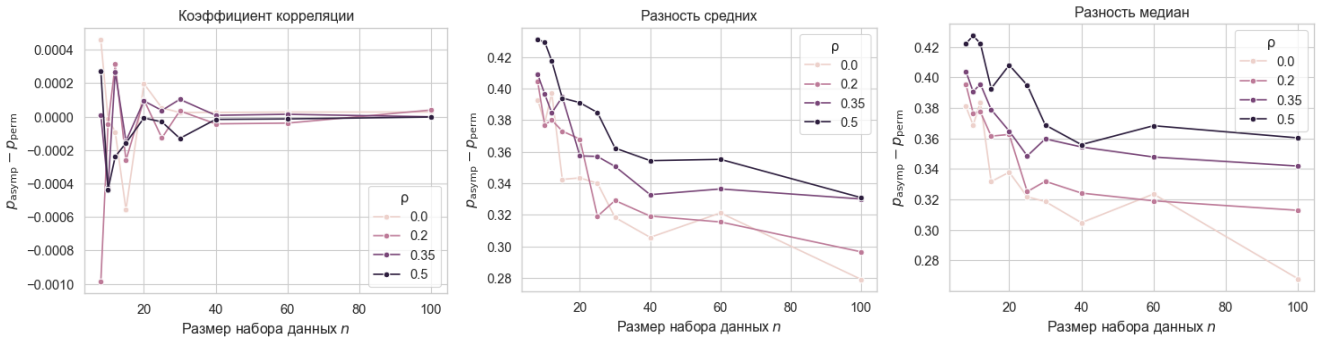


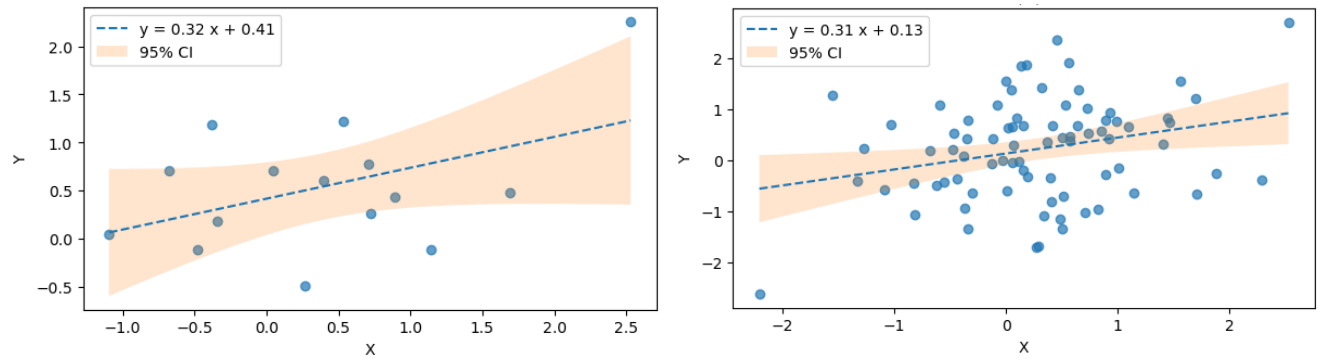
Рисунок 3.14 Зависимость расхождения между асимптотическим и пермутационным *p-value* от размера набора данных n

В результате численного эксперимента показано, что при малых объемах данных асимптотический *p-value* мог существенно отличаться от пермутационного *p-value*, при этом расхождение уменьшалось с увеличением числа наблюдений.

Для коэффициента корреляции видно, что при малых значениях n (менее 15–20 наблюдений) расхождения достигали максимальных значений, тогда как с ростом числа наблюдений расхождения уменьшались и при $n > 40$ приближались к нулю. В отличие от коэффициента корреляции, для разности средних и разности медиан наблюдается более медленное уменьшение расхождения, а сами значения отклонений остаются существенно выше нуля даже при числе наблюдений $n > 40$.

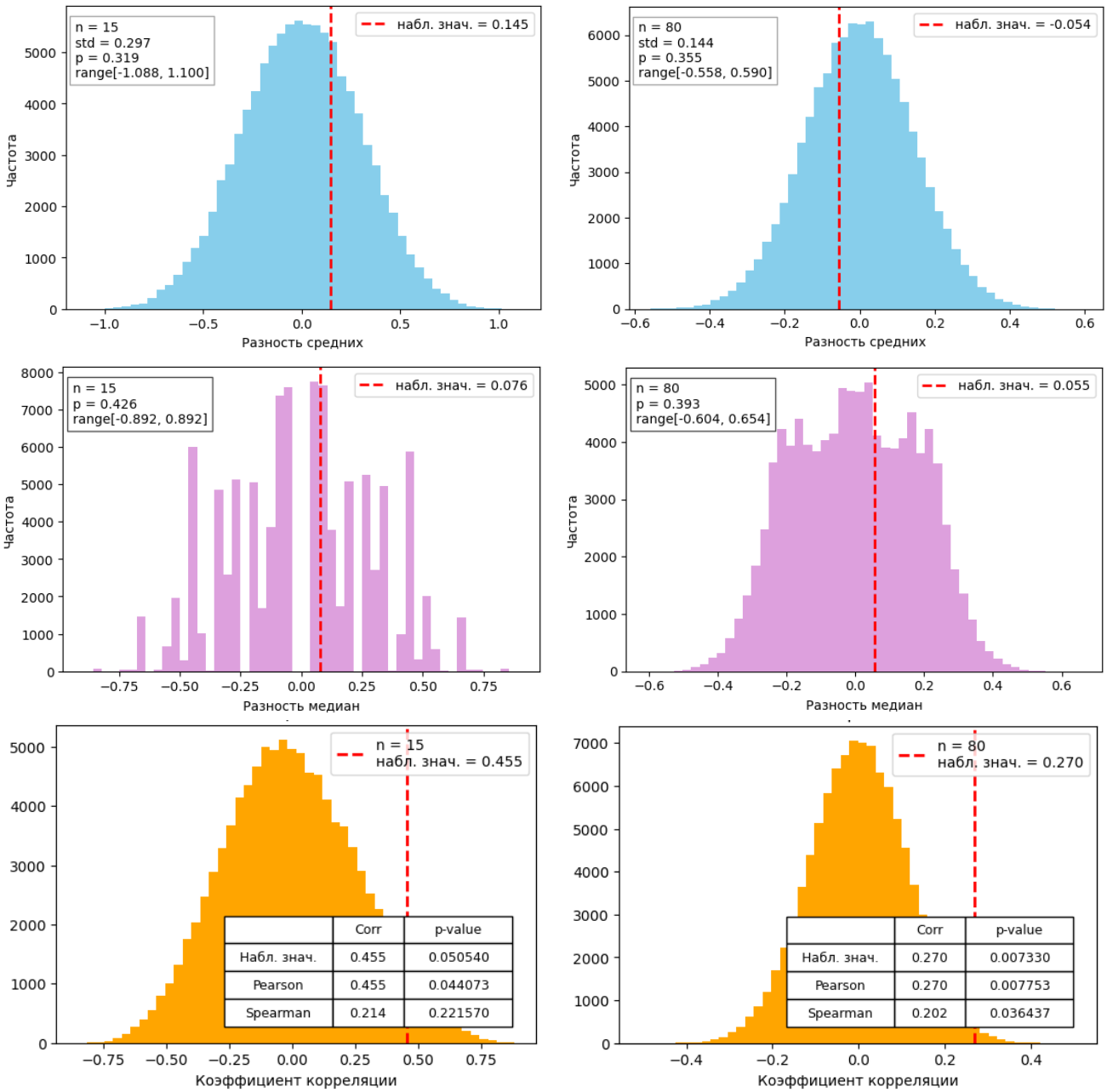
Для всех рассмотренных значений ρ наблюдается сходная динамика уменьшения расхождения между асимптотическими и пермутационными оценками *p-value*.

Для иллюстрации этого эффекта построены две конкретные реализации одной и той же модели при одинаковом значении $\rho = 0{,}35$, но при различном числе наблюдений ($n = 15$ и $n = 80$) (рисунки 3.15, 3.16).



$n = 15$ $n = 80$

Рисунок 3.15 Реализации модели $Y = \rho X + \sqrt{1-\rho^2}\varepsilon$ при $\rho = 0{,}35$ и числе наблюдений $n = 15$ и $n = 80$



$n = 15$ $n = 80$

Рисунок 3.16 Пример расчета перестановочных *p-value* для двух реализаций модели

На рисунке 3.16 приведены распределения трех выбранных статистик, полученные в результате перестановочного моделирования. Красной пунктирной линией отмечено наблюдаемое значение статистики. При малом объеме набора данных ($n = 15$) наблюдается более широкое распределение статистик, что отражает большую вариативность значений. При увеличении объема набора данных ($n = 80$) распределение становится более компактным, а наблюдаемое значение располагается ближе к центру распределения.

Средний ряд иллюстрирует распределения разности медиан. В отличие от разности средних, здесь сохраняется более выраженная дисперсия пермутационного распределения даже при увеличении числа наблюдений, что указывает на более медленную стабилизацию данной статистики относительно роста объема данных.

Нижний ряд представляет полученные распределения коэффициента корреляции, а также значения наблюдаемой корреляции и соответствующие оценки *p-value*, вычисленные асимптотическим и пермутационным подходами. Для случая $n = 15$ наблюдается заметное расхождение между *p-value* (уже во втором знаке после запятой), тогда как при $n = 80$ расхождение существенно уменьшается (в четвертом знаке после запятой).

Приведенный пример иллюстрирует, что при увеличении объема данных результаты асимптотического и пермутационного подходов становятся более согласованными.

Аналогичные результаты были получены для различных типов зависимостей между наборами наблюдаемых и предсказанных данных (Приложение Г).

На рисунках 3.17–3.19 приведены эмпирические распределения пермутационных статистик для модели Land Use Segmentation и моделей Land Use Buffers и Rings, показавших наилучшие значения RMSE, в численном эксперименте на синтетических данных при $\sigma = 0{,}2$. Для каждой статистики вычислялись соответствующие *p-value* на основе процедуры рандомизации ($N_{rand} = 10^6$).

В настоящей работе пермутационные *p-value* интерпретировались следующим образом. Для коэффициента корреляции низкие значения *p-value* указывали на наличие связи между наблюдаемыми и прогнозируемыми значениями. Соответственно, чем меньше значение *p-value*, тем выше качество модели и тем менее вероятно, что наблюдаемая корреляция получена случайно. Для разности средних и разности медиан, напротив, более высокие значения *p-value* интерпретировались как более высокая согласованность наблюдаемого и предсказанного распределений и, соответственно, более высокое качество модели.

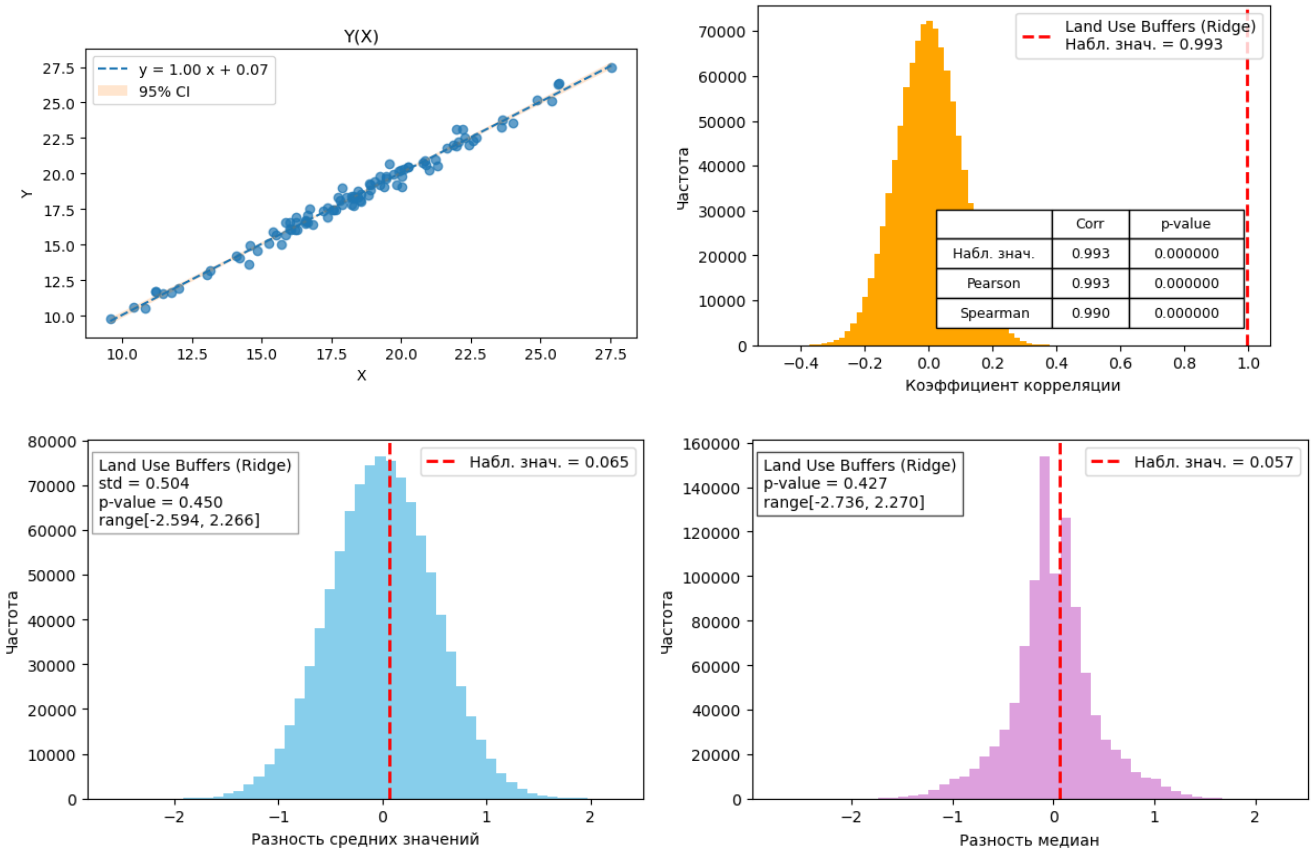


Рисунок 3.17 Оценка модели Land Use Buffers (Ridge) с помощью пермутационного подхода

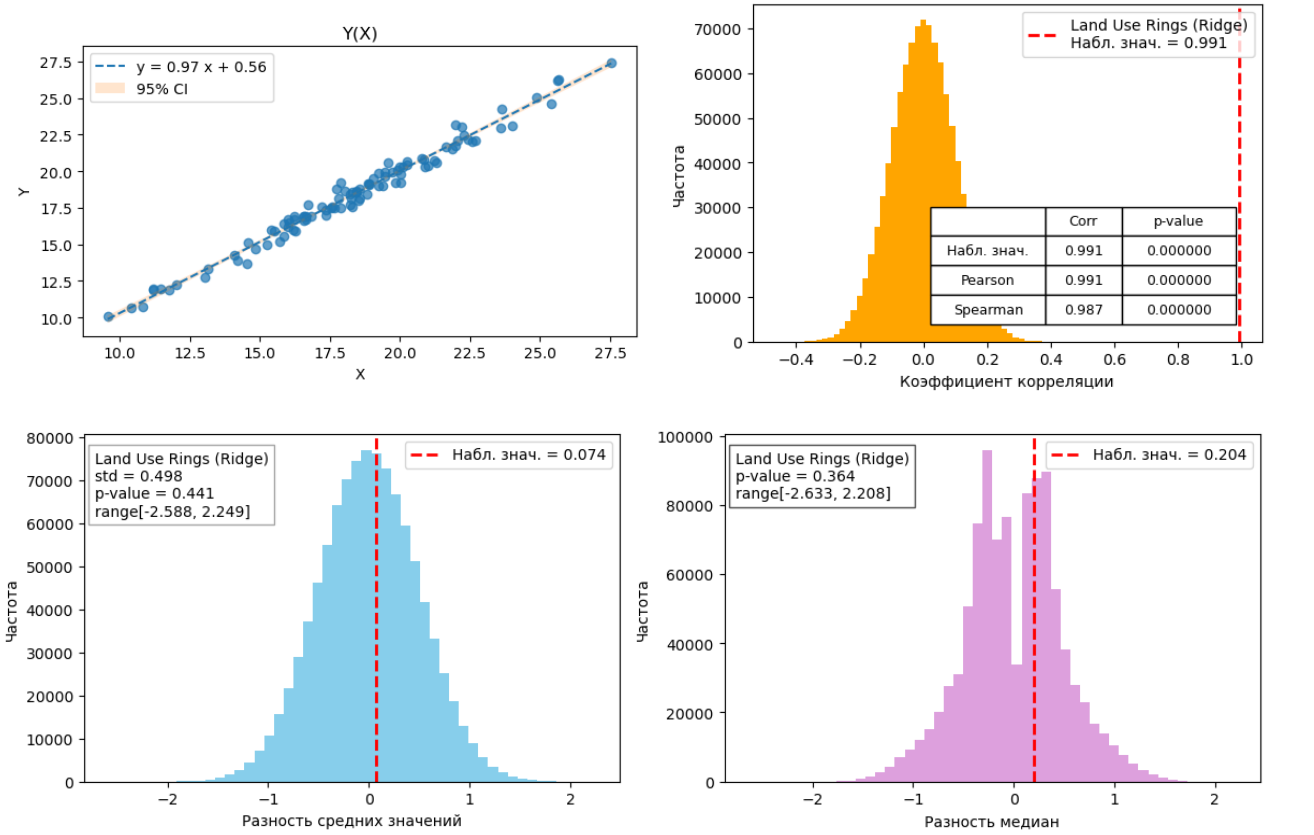


Рисунок 3.18 Оценка модели Land Use Rings (Ridge) с помощью пермутационного подхода

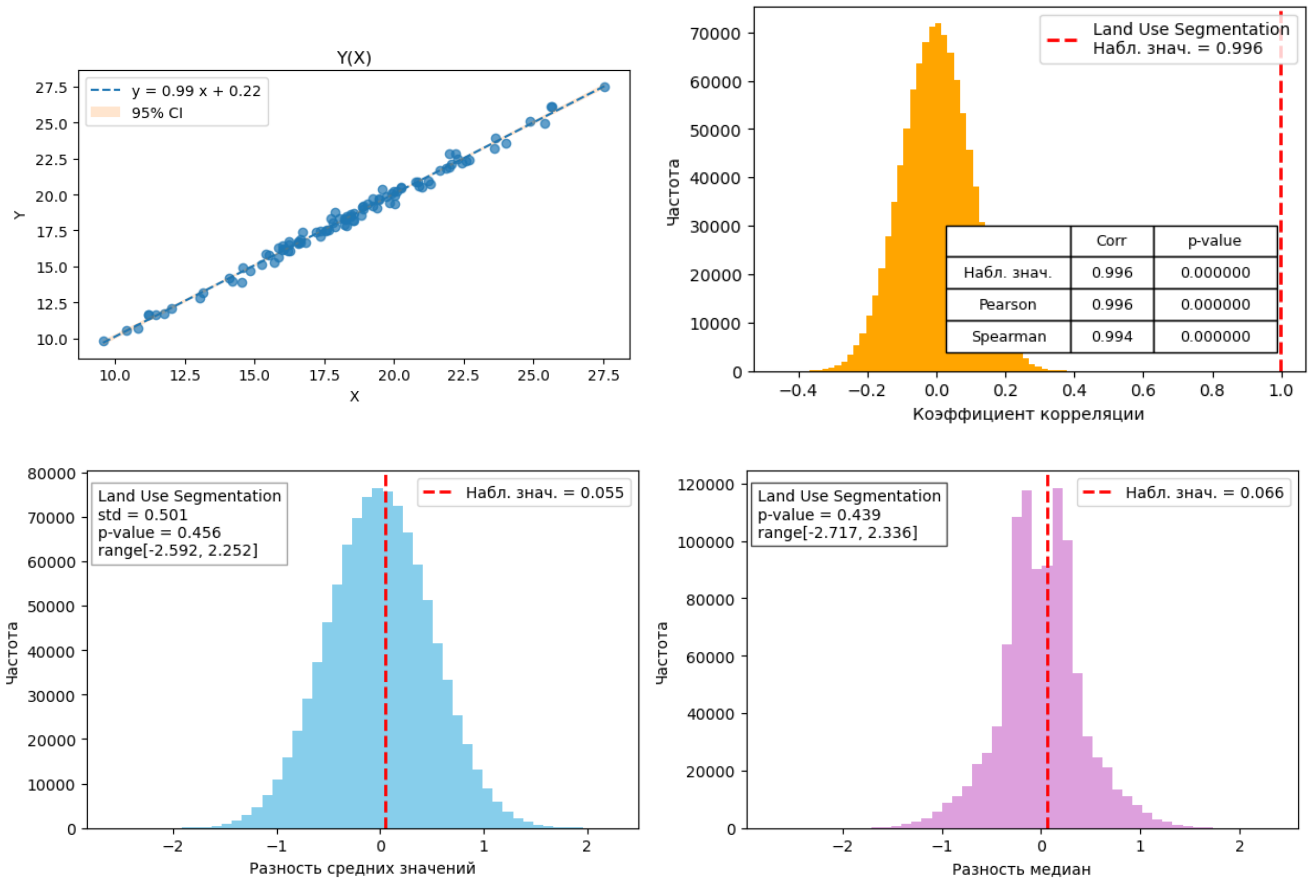


Рисунок 3.19 Оценка модели Land Use Segmentation с помощью пермутационного подхода

В таблице 3.7 показано, что для коэффициента корреляции *p-value* = 0,000 для всех построенных моделей. Таким образом, вероятность получения наблюдаемой корреляции случайным образом практически отсутствует. Однако поскольку значения *p-value* оказались одинаково низкими для всех моделей, данный показатель оказался малочувствительным к различиям между алгоритмами.

Для разности средних наиболее высокие значения *p-value* получены для моделей SVR: SVR (Rings): *p-value* = 0,498; SVR (Buffers): *p-value* = 0,496. Кроме того, высокие значения *p-value* были получены для Lasso (Buffers): *p-value* = 0,459, Land Use Segmentation: *p-value* = 0,456, Ridge (Buffers): *p-value* = 0,450, MLP (Rings): *p-value* = 0,443.

Для анализа разности медиан наилучшие результаты были получены для следующих моделей: SVR (Buffers): *p-value* = 0,486, Land Use Segmentation: *p-value* = 0,439, Ridge (Buffers): *p-value* = 0,427. Напротив, наиболее низкие значения p-value наблюдались для моделей kNN: kNN (Buffers): *p-value* = 0,048,

kNN (Rings): *p-value* = 0,002. Аналогичная тенденция, хотя и менее выраженная, наблюдалась для моделей Random Forest и Gradient Boosting.

3.3.3 Результаты численных экспериментов на данных наблюдений

В дополнение к традиционным метрикам, для оценки построенных в настоящей работе моделей на данных наблюдений потока пыли в снеговом покрове г. Новый Уренгой был применен пермутационный подход (таблица 3.8). На рисунках 3.20–3.22 приведены распределения пермутационных статистик для модели Land Use Segmentation и моделей Land Use Buffers и Rings, показавших наилучшие значения *RMSE*. Для каждой статистики вычислялись соответствующие *p-value* на основе процедуры рандомизации ($N_{rand} = 10^6$).

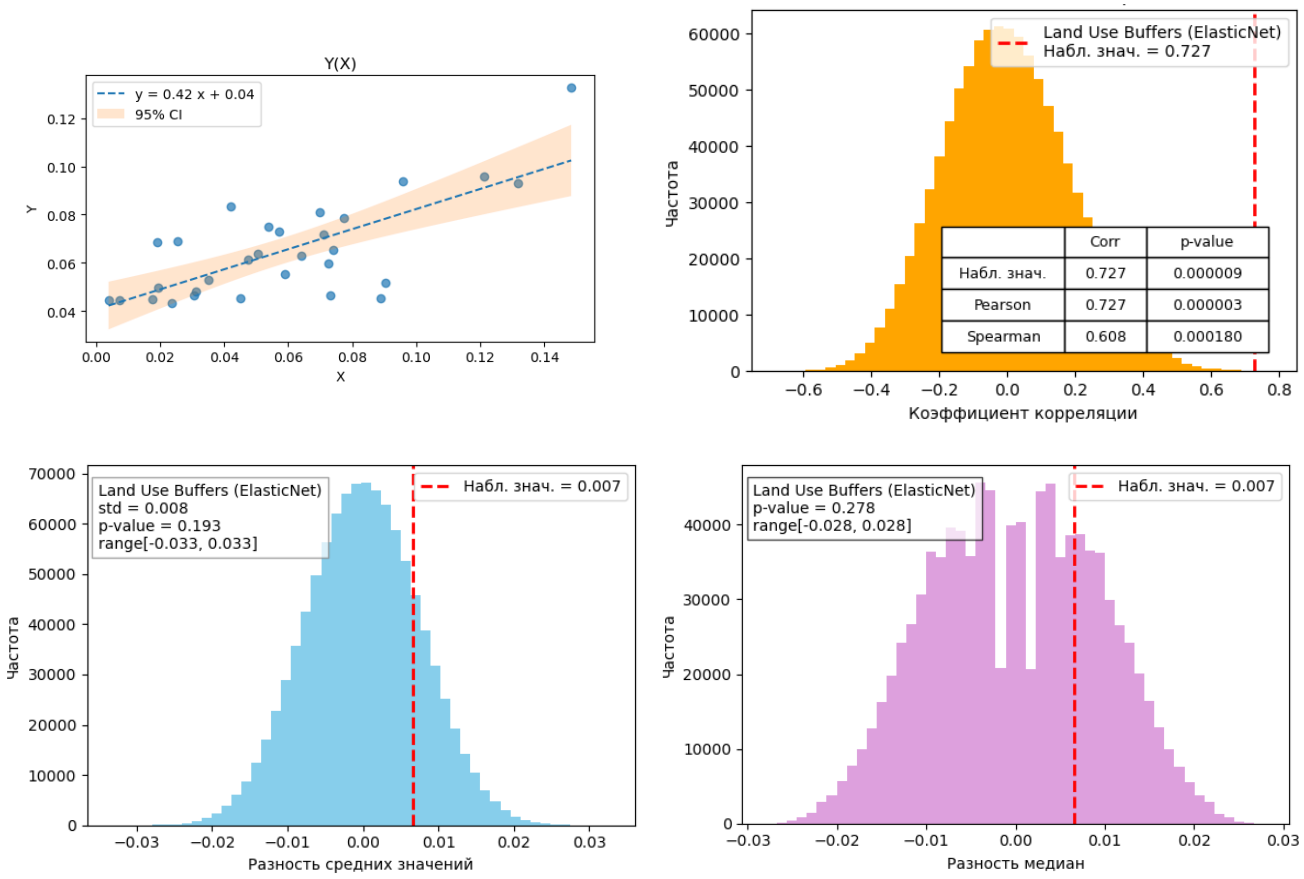


Рисунок 3.20 Оценка модели Land Use Buffers (ElasticNet) с помощью пермутационного подхода

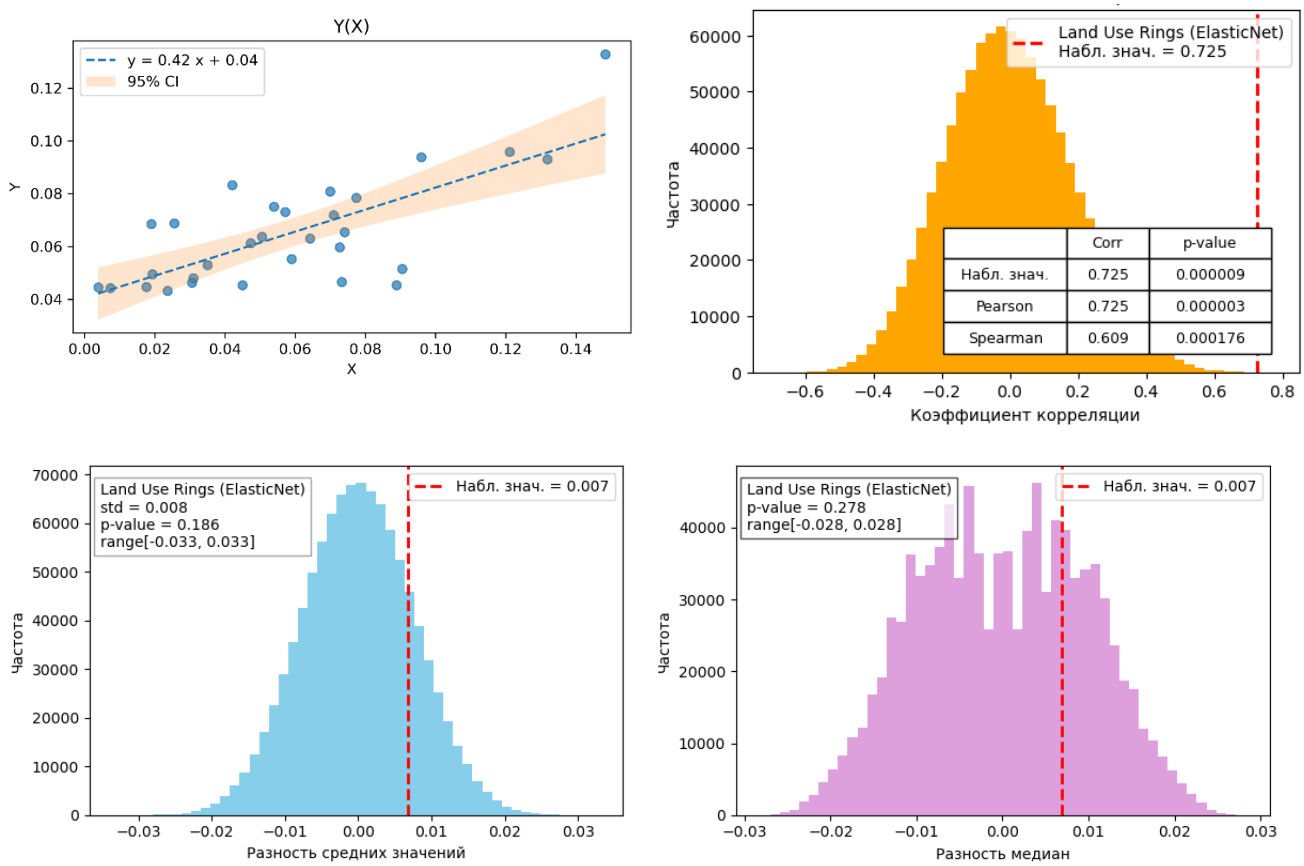


Рисунок 3.21 Оценка модели Land Use Rings (ElasticNet) с помощью пермутационного подхода

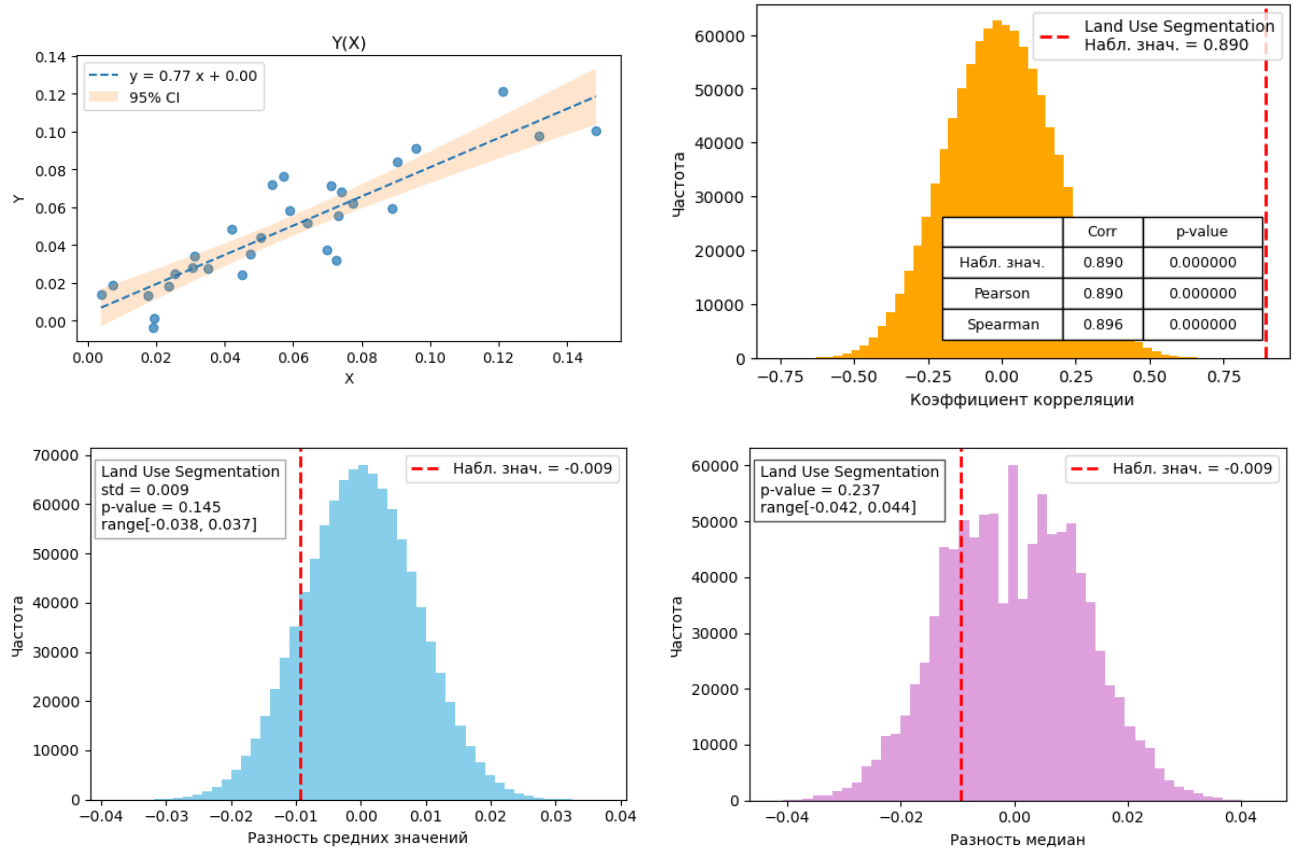


Рисунок 3.22 Оценка модели Land Use Segmentation с помощью пермутационного подхода

Для коэффициента корреляции наименьшие *p-value* < 0,001 были получены для модели Land Use Segmentation (*Corr* = 0,890) и моделей Land Use Buffers и Rings с аппроксиматорами ElasticNet и Rigde (ElasticNet: *Corr* = 0,727 для Buffers и *Corr* = 0,725 для Rings, Rigde: *Corr* = 0,704 для Buffers и *Corr* = 0,658 для Rings). Полученный результат свидетельствует о наличии связи между наблюдаемыми и предсказанными значениями.

Для разности средних значений наиболее высокие значения *p-value* наблюдались для моделей kNN (Land Use Buffers: *p-value* = 0,414) и RF (Land Use Buffers: *p-value* = 0,381; Land Use Rings: *p-value* = 0,316), а также GB (Land Use Rings: *p-value* = 0,362). Несмотря на относительно невысокие показатели R^2 и *Corr*, эти модели достаточно хорошо воспроизводят центральную тенденцию распределений.

Аналогичный результат получен для разности медиан. Наиболее высокие значения *p-value* получены для RF (Land Use Buffers: *p-value* = 0,500; Land Use Rings: *p-value* = 0,488) и kNN (Land Use Buffers: *p-value* = 0,497). Это свидетельствует о высокой согласованности медианных характеристик наблюдаемого и предсказанного наборов данных.

При этом модель Land Use Segmentation, демонстрирующая лучшие значения традиционных метрик качества (R^2 = 0,723, *RMSE* = 0,019), характеризуется сравнительно более низкими значениями *p-value* для разности средних (*p-value* = 0,145) и разности медиан (*p-value* = 0,237). Такой результат указывает на наличие небольшого систематического смещения распределения относительно наблюдаемых данных. Этот эффект также подтверждается отрицательными значениями $Diff_{means}$ = –0,009 и $Diff_{medians}$ = –0,009, свидетельствующими о некоторой тенденции модели к занижению прогнозируемых значений.

Таблицы 3.7–3.12 иллюстрируют, что пермутационный подход может рассматриваться как дополнение к традиционным метрикам при оценке качества моделей пространственного распределения примесей на малых наборах данных.

Пермутационный подход был также применен для оценки трех ИНС-моделей (MLP, RBF и GRNN), прогнозирующих концентрации Cu и Fe в почвах

субарктических городов [3]. В качестве статистик использовались разность средних и коэффициент корреляции. Традиционные индексы практически однозначно указывали на MLP как наилучшую модель, в то время как пермутационные *p-value* продемонстрировали дополнительные различия между моделями. Так, пермутационные *p-value* показали, что RBF и GRNN статистически почти не уступают или даже превосходят MLP, указывая на меньшую уверенность в явном превосходстве этой модели. Примеры таких результатов:

- Новый Уренгой, Cu: *p-value* (*Corr*) = 0,01–0,02 для RBF и GRNN против *p-value* (*Corr*) = 0,03 для MLP;

- Новый Уренгой, Fe: *p-value* ($Diff_{means}$) = 0,17–0,20 для GRNN и RBF против *p-value* ($Diff_{means}$) = 0,10 для MLP;

- Ноябрьск, Cu: *p-value* (*Corr*) = 0,03 для GRNN против *p-value* (*Corr*) = 0,16 для MLP;

- Ноябрьск, Fe: *p-value* ($Diff_{means}$) = 0,21–0,24 для RBF и GRNN против *p-value* ($Diff_{means}$) = 0,09 для MLP.

Преимущества пермутационного подхода продемонстрированы на временных данных динамики концентрации метана в приземном слое атмосферы о. Белый при построении моделей, сочетающих ИНС Elman и NARX с предварительной дискретной вейвлет-декомпозицией наблюдаемой реализации временного ряда на основе вейвлетов Daubechies (db4) и Symlet (sym4) [2]. В ряде ситуаций пермутационные *p-value* дополняли или корректировали выводы, основанные на традиционных метриках. Например, хотя модель NARX (sym4) выглядела наилучшей по значениям *MAE*, *RMSE* и коэффициента корреляции, пермутационное *p-value* для разности средних показало, что только модель Elman (sym4) имеет статистически незначимое отклонение предсказанных значений от наблюдаемых (*p-value* = 0,178). Аналогично, несмотря на лидерство модели NARX (db4) по индексам согласия, пермутационные *p-value* для корреляции вновь указали на то, что Elman (sym4) демонстрирует наиболее статистически надежную согласованность с наблюдаемым рядом.

Таким образом, во всех рассмотренных случаях пермутационный подход показал, насколько статистически значимы различия между наблюдением и прогнозом, и обеспечил более полную и объективную оценку качества моделей по сравнению с использованием исключительно традиционных метрик.

3.4 Оценка репрезентативности точек наблюдений

Раздел посвящен исследованию репрезентативности точек наблюдений в задаче пространственного моделирования. Рассматривается влияние точек набора данных на качество построения моделей.

3.4.1 Постановка численного эксперимента

Использовалось сгенерированное пространственное распределение, в котором заранее задана репрезентативная область (hotspot). Точки наблюдений были независимо и равномерно распределены на единичном квадрате:

$$x_i, y_i \sim \mathcal{U}([0,1]), i = 1, \dots, n.$$

Наблюдаемая целевая переменная z_i формировалась как сумма трех компонентов:

- фон (background) – линейная пространственная трендовая компонента:

$$b(x_i, y_i) = 0{,}5x_i + 0{,}3y_i,$$

- локализованная информативная структура (hotspot) – гауссовский пик, моделирующий репрезентативную область:

$$h(x_i, y_i) = 5 \cdot \exp\left(\quad -\frac{(x_i - 0{,}3)^2 + (y_i - 0{,}7)^2}{0{,}01} \right),$$

- стохастический шум наблюдений:

$$\varepsilon_i \sim \mathcal{N}(0, 0{,}2^2).$$

Таким образом, наблюдение задавалось как:

$$s_i = b(x_i, y_i) + h(x_i, y_i) + \varepsilon_i.$$

Для задания репрезентативности точки вводилась бинарная индикаторная функция принадлежности к области hotspot. Точка считалась репрезентативной, если она попадала в окрестность центра гауссовской структуры.

Таким образом, синтетический набор данных представлял собой комбинацию репрезентативных и фоновых наблюдений.

3.4.2 Результаты численных экспериментов на синтетических данных

На синтетической территории показано сравнение истинной репрезентативности с результатами алгоритма выделения репрезентативных точек (рисунок 3.23). Все сгенерированные точки равномерно распределены по двумерному пространству. На этом фоне выделяется локализованная область репрезентативности (hotspot). Данная область соответствует зоне, в которой вклад нелинейной компоненты целевой функции максимален и, следовательно, наблюдения обладают наибольшей информационной ценностью для восстановления пространственной структуры.

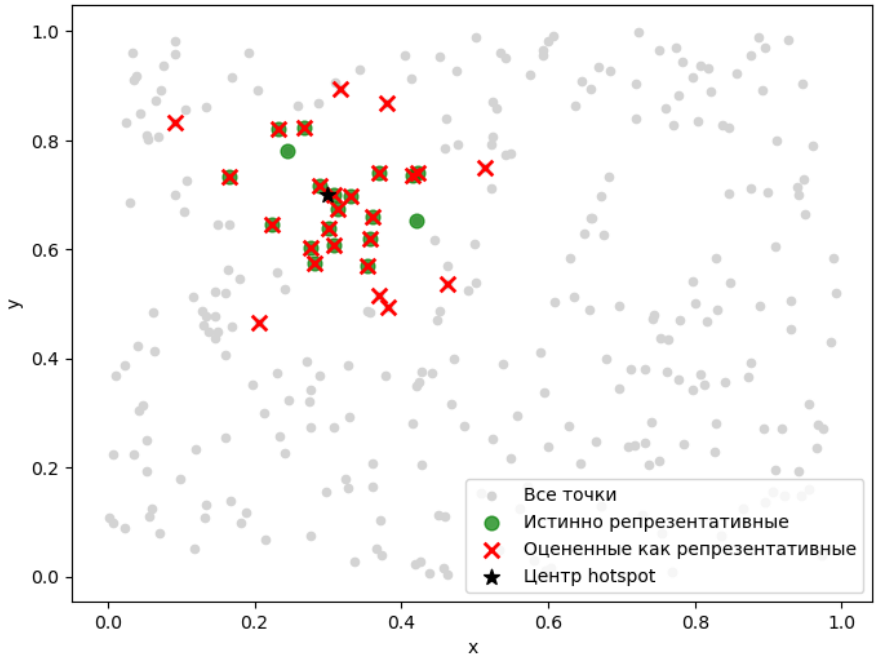


Рисунок 3.23 Оценка репрезентативности точек наблюдений на синтетических данных

Результаты оценки репрезентативности наблюдений представлены отдельным множеством точек. Эти точки преимущественно концентрируются внутри области hotspot, однако также включают небольшое число ложноположительных срабатываний, расположенных вблизи границ зоны высокой значимости.

Таким образом, наблюдается высокая степень совпадения между оцененными и истинно репрезентативными точками с ограниченным количеством ошибок классификации, выражающихся в отдельных ложноположительных точках вблизи области перехода. Полученный результат свидетельствует о том, что метод оценки репрезентативности способен выявлять репрезентативные пространственно локализованные области в условиях контролируемого синтетического эксперимента.

3.4.3 Результаты численных экспериментов на данных наблюдений

Численные эксперименты по оценке репрезентативности точек наблюдений были выполнены для данных содержания элементов в почвах городов Ноябрьск и Тарко-Сале. Для каждого исследуемого элемента проводилась серия многократных случайных разбиений исходного набора данных на обучающее и тестовое подмножества, построение модели многослойного персептрона и вычисление метрик качества прогноза.

Оценка репрезентативности точек основывалась на анализе их участия в обучающих подмножествах, обеспечивающих наилучшие показатели качества модели. Для каждой точки рассчитывалась частота ее попадания в обучающие подмножества моделей, относящихся к группе лучших реализаций. Полученная величина интерпретировалась как характеристика вклада данной точки в формирование моделей с низкими ошибками прогноза.

Для оценки распределения репрезентативности по точкам наблюдения были построены кривые, отражающие зависимость числа попаданий точки в лучшие обучающие подмножества от ее ранга. Ранг определялся путем сортировки точек по убыванию значения *RMSE* модели (рисунок 3.24).

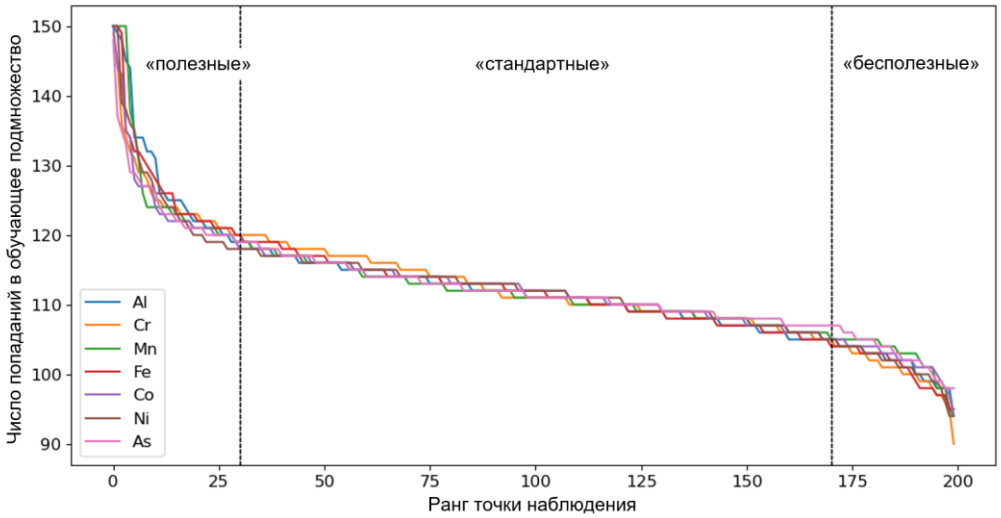


Рисунок 3.24 Оценка репрезентативности для исследуемых элементов на территории г. Ноябрьск

Анализ полученных зависимостей показывает наличие неоднородного вклада отдельных точек в процесс построения моделей. Для большинства исследуемых элементов наблюдается характерная форма кривой с выраженными тремя областями. В левой части графика располагаются точки, характеризующиеся высокой частотой попадания в лучшие обучающие подмножества. Такие точки можно интерпретировать как наиболее репрезентативные (или «полезные»), поскольку их присутствие в обучающем подмножестве связано с улучшением качества прогноза. Центральная область кривой соответствует точкам со средней уровнем репрезентативности («стандартные» точки). Их вклад в качество модели является менее выраженным. Правая часть кривой содержит точки с минимальной частотой попадания в лучшие обучающие подмножества. Такие точки оказывают незначительное влияние на качество прогнозирования и могут рассматриваться как малорепрезентативные (или «бесполезные» в смысле их ценности для обучения модели). Полученные результаты указывают на то, что исходный набор данных обладает неоднородной структурой с точки зрения информативности отдельных наблюдений.

Для исследования пространственных закономерностей распределения репрезентативных наблюдений выполнялась их визуализация на карте исследуемой территории (рисунок 3.25). Для каждого химического элемента

выделялось подмножество точек с максимальными значениями показателя репрезентативности. Дополнительно проводился анализ пространственного пересечения выделенных множеств для различных элементов.

Анализ пространственного расположения показывает, что наиболее репрезентативные точки распределены по территории неравномерно. Для ряда элементов наблюдается концентрация таких точек в областях с выраженными пространственными особенностями, включая границы исследуемой территории, участки с повышенной изменчивостью концентраций и зоны с возможным влиянием локальных источников загрязнения.

Кроме того, для различных элементов наблюдается частичное пространственное перекрытие множеств репрезентативных точек. Наличие совпадающих точек может указывать на существование общих пространственных закономерностей формирования распределений концентраций исследуемых элементов. В то же время наличие уникальных для отдельных элементов точек свидетельствует о специфике механизмов формирования их пространственной структуры.

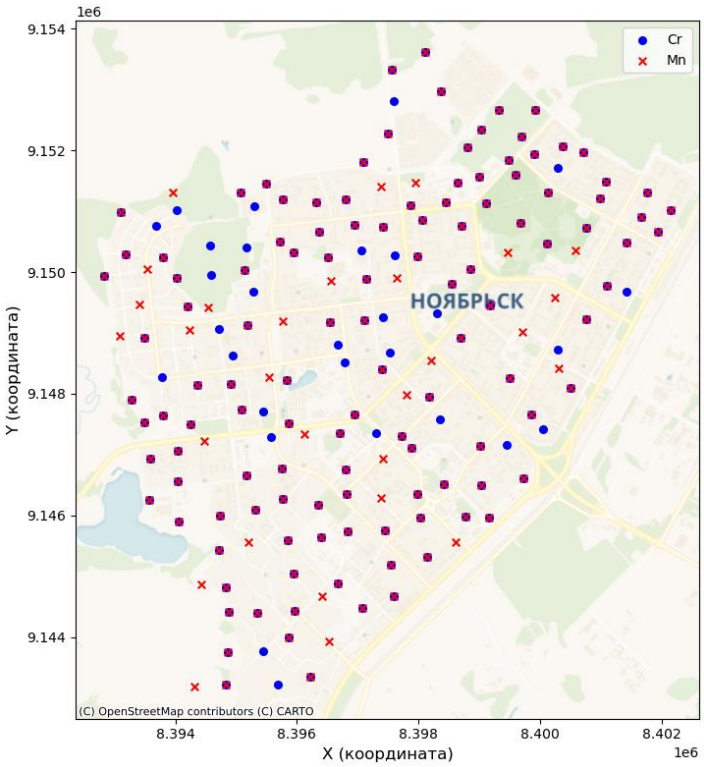


Рисунок 3.25 Пространственное распределение репрезентативных точек для Cr и Mn на территории г. Ноябрьск

Аналогичная процедура оценки репрезентативности была выполнена для данных г. Тарко-Сале (рисунок 3.26). Отсортированные по репрезентативности точки для четырех металлов позволили выделить три класса: «полезные», «стандартные» и «бесполезные» для обучения модели точки наблюдений.

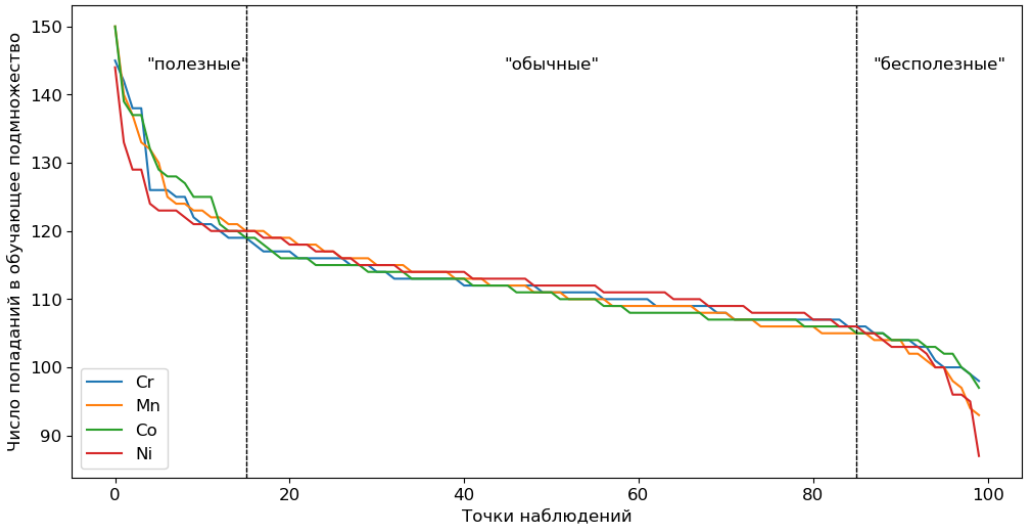


Рисунок 3.26 Оценка репрезентативности для исследуемых элементов на территории г. Тарко-Сале

Пространственное распределение наиболее репрезентативных наблюдений выполнено на карте исследуемой территории (рисунок 3.27).

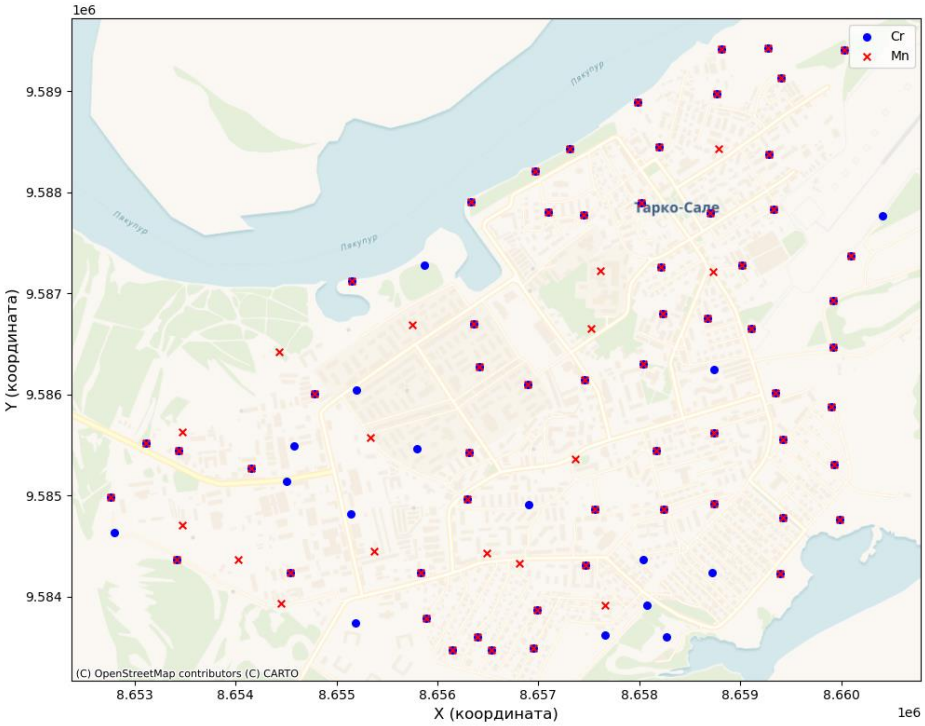


Рисунок 3.27 Пространственное распределение репрезентативных точек для Cr и Mn на территории г. Тарко-Сале

Сравнение результатов для двух исследуемых территорий показывает, что структура распределения репрезентативных точек зависит от пространственного расположения наблюдений. При этом общая форма кривых репрезентативности сохраняется и характеризуется наличием ограниченного числа высокоинформативных наблюдений, формирующих основную часть обучающих подмножеств, участвующих в обучении «лучших» моделей. В итоговое обучающее подмножество чаще всего попадали граничные точки и точки, отражающие территориальные характеристики (морфология, структура улиц и др.). Поскольку элементный состав являлся однородным и без выраженных аномалий, все точки имели практически равные шансы быть включенными в обучение.

Оценка индивидуальной репрезентативности точек наблюдений по данным о содержании хрома (Cr) и марганца (Mn) в верхнем слое почвы в г. Ноябрьск представлена в работе [15].

3.5 Обсуждение результатов

В настоящей работе показано влияние структуры пространственных предикторов на устойчивость оценок коэффициентов моделей Land Use Buffers и Land Use Rings. Предложен переход от вложенных круговых окрестностей к непересекающимся кольцевым как способ разрешения одного из ограничений классического подхода Land Use – мультиколлинеарности предикторов на основе вложенных круговых окрестностей точек наблюдений. Проведенные численные эксперименты на синтетических данных и данных наблюдений показали, что использование круговых окрестностей приводит к систематическому перекрытию областей агрегации пространственных характеристик территории, что вызывает выраженную линейную зависимость между предикторами. Это проявляется в высоких значениях парных корреляций, существенном росте фактора инфляции дисперсии и значениях числа обусловленности, соответствующих плохо обусловленным регрессионным задачам. Переход к кольцевым пространственным переменным приводил к снижению мультиколлинеарности и улучшению обусловленности матрицы предикторов.

Полученные результаты позволяют сформулировать рекомендации по построению пространственных переменных-предикторов на основе окрестностей точек наблюдений:

1. Предпочтительно использовать непересекающиеся кольцевые окрестности вместо вложенных круговых.

2. В ситуациях, когда использование вложенных круговых окрестностей необходимо для задач исследователя (например, для анализа суммарного влияния всех источников в пределах заданного радиуса) рекомендуется анализировать структуру матрицы пространственных переменных и использовать методы пошагового отбора предикторов или регуляризации (Ridge, LASSO).

3. При анализе структуры матрицы пространственных предикторов рекомендуется выполнять центрирование и стандартизацию предикторов, анализировать парные корреляции между предикторами, рассчитывать фактор инфляции дисперсии, оценивать число обусловленности матрицы признаков.

4. Если после отбора переменных матрица сохраняет высокую обусловленность, рекомендуется использовать альтернативные способы формирования пространственных переменных: переход к кольцевым окрестностям, укрупнение пространственных диапазонов.

При этом анализ результатов численных экспериментов показал, что высокая точность прогноза не является достаточным условием устойчивости модели. Несмотря на сопоставимые значения метрик качества, модели, построенные на различных наборах пространственных предикторов, демонстрируют существенно различающуюся чувствительность к изменению обучающего подмножества и структуре входных данных. Полученные результаты подтверждают, что минимизация функции ошибки не гарантирует надежность модели в задачах пространственного моделирования, особенно при ограниченном объеме и неоднородности наблюдений. Это обосновывает необходимость совместного учета показателей точности и устойчивости при сравнительной оценке моделей.

Модель Land Use Segmentation показала наилучшее качество по сравнению с моделями Land Use Buffers и Land Use Rings: это проявлялось как в более высоких

значениях коэффициента корреляции и коэффициента детерминации, так и в снижении ошибки прогнозирования. Этот результат указывает на более корректное восстановление пространственного распределения примеси и подтверждает преимущество модели Land Use Segmentation при работе в условиях дефицита наблюдений, характерного для задач экологического мониторинга. В целом результаты пермутационного подхода согласуются с традиционными метриками качества. Модели, характеризующиеся высокими значениями коэффициентов корреляции и детерминации и низкими значениями ошибок, как правило, демонстрировали и более высокую согласованность распределений. Однако полученные результаты подтверждают, что пермутационные *p-value* позволяют выявлять расхождения между моделями, которые не видны по традиционным метрикам качества, и тем самым уточняют и дополняют процедуру оценки моделей.

Проведенные численные эксперименты показали, что использование традиционных асимптотических критериев может приводить к нестабильным и потенциально смещенным выводам о качестве моделей, особенно при малых наборах данных и сложной структуре пространственной зависимости. В этих условиях пермутационный подход обеспечивает более корректную оценку значимости за счет отказа от предположений о распределении ошибок и моделирования распределения статистики напрямую. Таким образом можно сформулировать еще одну рекомендацию: всегда использовать пермутационный подход в задаче оценки моделей пространственного распределения примесей.

Дополнительно установлено, что различные группы моделей по-разному воспроизводят структуру исходных данных. Кригинг позволил восстановить общую пространственную структуру поля, однако не объяснил механизмы формирования локальных зон повышенных содержаний примеси. Модели Land Use позволили восстановить пространственное распределение примеси с учетом характеристик источников, демонстрируя согласованность получаемых распределений с конфигурацией источников примеси на территории. Наилучшие результаты продемонстрировала модель Land Use Segmentation, для которой

характерны минимальные ошибки прогнозирования и более гладкая пространственная структура при сохранении основных закономерностей распределения примеси. Модели Land Use Buffers и Land Use Rings также корректно воспроизводили общую структуру пространственного поля, однако характеризовались большей локальной вариабельностью и менее сглаженным характером распределений. Сравнение линейных и нелинейных моделей показало, что нелинейные методы формируют более неоднородные пространственные поля и сильнее выделяют локальные зоны накопления примеси.

Наконец, было показано, что точки наблюдений обладают неравной репрезентативностью в задаче моделирования пространственного распределения в смысле их ценности для обучения модели. Отсортированные по репрезентативности точки для двух металлов позволили выделить три класса: «полезные», «стандартные» и «бесполезные» для обучения модели точки наблюдений. При этом традиционные представления о репрезентативности выборки, основанные исключительно на пространственном покрытии территории, не в полной мере отражают способность отдельных наблюдений описывать структуру моделируемого поля.

3.6 Выводы по главе 3

1. В численных экспериментах показано, что структура пространственных переменных оказывает влияние на устойчивость оценок коэффициентов моделей Land Use. Использование вложенных круговых окрестностей (Land Use Buffers) приводит к мультиколлинеарности, обусловленной накопительным характером предикторов, тогда как предложенный подход Land Use Rings на основе непересекающихся кольцевых окрестностей устраняет дублирование информации между предикторами. В серии численных экспериментов Land Use Rings снижал мультиколлинеарность предикторов, измеренную по фактору инфляции дисперсии, в среднем в 19,16 раз и повышал устойчивость оценок коэффициентов регрессионной модели, измеренную по числу обусловленности матрицы $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, в 7,52 раза по сравнению с классическим подходом Land Use Buffers.

2. Численные эксперименты показали, что модели Land Use Buffers и Land Use Rings обеспечивают сопоставимое качество моделирования пространственного распределения примесей, при этом наилучшие результаты для рассмотренного синтетического сценария и данных наблюдений потока пыли в снеговом покрове достигались при использовании регуляризованных линейных методов (Ridge, ElasticNet). Различия между Buffers и Rings по метрикам качества в большинстве случаев невелики, однако переход к кольцевым предикторам позволяет существенно повысить устойчивость моделей за счет снижения мультиколлинеарности без потери точности. При малых объемах наблюдений и наличии шума регуляризованные линейные модели оказались более надежными по сравнению с ансамблевыми и нейросетевыми подходами.

3. Результаты численных экспериментов на синтетических данных и данных наблюдений показали, что предложенная модель Land Use Segmentation обеспечивает более высокое качество моделирования пространственного распределения примеси по сравнению с подходами Land Use Buffers и Land Use Rings. Для Land Use Segmentation получены более высокие значения коэффициентов детерминации и корреляции, а также меньшие значения ошибок *MAE* и *RMSE*. Дополнительно показано, что Land Use Segmentation характеризуется большей устойчивостью к шуму измерений и формирует более гладкие и физически согласованные пространственные распределения, воспроизводящие структуру наблюдаемого поля примеси.

4. В результате численных экспериментов показано, что при малых размерах набора данных асимптотические *p-value* могут существенно расходиться с пермутационными *p-value* и приводить к некорректным выводам о качестве модели. Пермутационный подход к оценке моделей, который является точным, не требует предположений о виде распределений и допускает использование любой подходящей статистики. Предложенный пермутационный подход к оценке качества моделей дополнил традиционный набор метрик качества вероятностной интерпретацией согласия/расхождения между предсказанными и наблюдаемыми данными. Использование пермутационного подхода выявило ряд ситуаций, в

которых пермутационные *p-value* дополняли или корректировали выводы, основанные на стандартных показателях точности. Эти результаты показывают, что пермутационная оценка может изменить ранжирование моделей и обеспечить более обоснованный выбор наилучшей прогностической модели.

5. Показано, что точки наблюдений обладают неравной репрезентативностью, оцениваемой как относительная частота попадания этих точек в обучающие подмножества, соответствующие таким разбиениям набора точек наблюдений на обучающее и тестовое подмножества, для которых достигаются наилучшие показатели *RMSE* модели. Для двух исследуемых территорий построена зависимость частоты попадания точки в наилучшие обучающие подмножества (по *RMSE* модели), в результате точки наблюдений сформировали три класса: высокорепрезентативные («полезные»), формирующие обучающие подмножества, связанные с минимальными значениями ошибки, точки средней информативности («стандартные») и малорепрезентативные («бесполезные») наблюдения с низкой частотой участия в «лучших» разбиениях.

Результаты третьей главы опубликованы в работах [1–7, 9, 10, 14–17].

Заключение

В результате проведенных исследований можно сделать следующие выводы:

1. Анализ существующих методов и подходов к построению математических моделей пространственного распределения примесей в компонентах окружающей среды показал, что как физические модели, так и модели аппроксимации демонстрируют ограничения при работе с малыми и разреженными наборами данных измерений. Обоснована целесообразность использования и развития подхода Land Use как этапа построения пространственных переменных-предикторов на основе доступных данных о землепользовании.

2. Разработан подход к построению пространственных переменных Land Use Rings, снижающий мультиколлинеарность между предикторами модели. Реализованы соответствующие численные методы построения пространственных переменных и проведен численный анализ устойчивости оценок коэффициентов регрессионных моделей, предикторы для которых получены с применением разработанного подхода Land Use Rings и классического подхода Land Use Buffers. Показано, что применение разработанного подхода повышает устойчивость оценок коэффициентов регрессионной модели по сравнению с Land Use Buffers.

3. Разработана математическая модель Land Use Segmentation и соответствующие ей численные методы и показано, что модель снижает ошибку аппроксимации пространственного распределения примесей в условиях ограниченных данных наблюдений. Построены модели на основе линейной регрессии, ансамблевых алгоритмов, метода опорных векторов, метода k-ближайших соседей и искусственной нейронной сети, использующие предикторы на основе круговых и кольцевых окрестностей точек наблюдений Land Use Rings и Land Use Buffers. Разработанные модели и методы реализованы в виде программного комплекса для моделирования пространственного распределения примесей.

4. Разработан метод оценки репрезентативности точек наблюдений и выполнено его применение для анализа информативности различных конфигураций точек наблюдений в задаче построения пространственного

распределения примеси. Оценка индивидуальной или групповой репрезентативности точек наблюдений позволяет в перспективе сократить объем полевых работ без существенной потери точности модели.

5. Проведено тестирование разработанных моделей на синтетических данных и их апробация на данных измерений примесей в различных компонентах окружающей среды. Выполнено сравнение с моделями, использующими классический подход к построению пространственных переменных Land Use Buffers. Апробирован предложенный пермутационный подход к оценке моделей. Показано, что модель Land Use Segmentation демонстрирует наилучшее качество, оцененное как с использованием стандартных метрик, так и с помощью пермутационного подхода.

Таким образом, поставленная в работе цель была достигнута, задачи решены в полном объеме. Направления дальнейших исследований могут включать развитие подходов к построению пространственных переменных и моделей Land Use с учетом анизотропии переноса примесей, в том числе расширение класса физически мотивированных моделей и дополнение моделей новыми данными о землепользовании, исследование переноса разработанных подходов и моделей на другие типы примесей и компоненты окружающей среды, а также на территории с различной пространственной структурой и климатическими условиями.

Список использованных источников

1. Методы расчетов рассеивания выбросов вредных (загрязняющих) веществ в атмосферном воздухе : утв. приказом Минприроды России от 06.06.2017 № 273. — М. : Минприроды России, 2017.

2. ETC/ATNI. (2021). Health risk assessments of air pollution: Estimations of the 2019 HRA, benefit analysis of reaching specific air quality standards and more (ETC/ATNI Report 2021/10). European Topic Centre on Air pollution, transport, noise and industrial pollution.

3. The spatial–temporal effect of air pollution on individuals' reported health and its variation by ethnic groups in the United Kingdom: a multilevel longitudinal analysis / Abed Al Ahad M., Demšar U., Sullivan F., and Kulu H. // *BMC Public Health*. — 2023. — Vol. 23, no. 1. — P. 897.

4. Opportunities and challenges for filling the air quality data gap in low-and middle-income countries / Pinder R. W., Klopp J. M., Kleiman G., Hagler G. S., Awe Y., and Terry S. // *Atmospheric Environment*. — 2019. — Vol. 215. — P. 116794.

5. Jemeļjanova M., Kmoch A., Uuemaa E. Adapting machine learning for environmental spatial data: a review // *Ecological Informatics*. — 2024. — Vol. 81. — P. 102634.

6. Environmental Monitoring for Arctic Resiliency and Sustainability: An Integrated Approach with Topic Modeling and Network Analysis / Zhu X., Pasch T. J., Ahajjam M. A., and Bergstrom A. // *Sustainability*. — 2022. — Vol. 14, no. 24. — P. 16493.

7. Emerging technologies and approaches for in situ, autonomous observing in the Arctic / Lee C. M., DeGrandpre M., Guthrie J., Hill V., Kwok R., Morison J., Cox C. J., Singh H., Stanton T. P., and Wilkinson J. // *Oceanography*. — 2022. — Vol. 35, no. 3–4. — P. 210–221.

8. Mapping urban air pollution using GIS: a regression-based approach / Briggs D. J., Collins S., Elliott P., Fischer P., Kingham S., Lebret E., Pryl K., van Reeuwijk H., Smallbone K., and van der Veen A. // *International Journal of Geographical Information Science*. — 1997. — Vol. 11, no. 7. — P. 699–718.

9. A review of land-use regression models to assess spatial variation of outdoor air pollution / Hoek G., Beelen R., De Hoogh K., Vienneau D., Gulliver J., Fischer P., and Briggs D. // *Atmospheric Environment*. — 2008. — Vol. 42, no. 33. — P. 7561–7578.

10. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide / Chen J., de Hoogh K., Gulliver J., Hoffmann B., Hertel O., Ketzel M., ... and Hoek G. // *Environment International*. — 2019. — Vol. 130. — P. 104934.

11. Incorporating land-use regression into machine learning algorithms in estimating the spatial-temporal variation of carbon monoxide in Taiwan / Wong P. Y., Hsu C. Y., Wu J. Y., Teo T. A., Huang J. W., Guo H. R., ... and Spengler J. D. // *Environmental Modelling & Software*. — 2021. — Vol. 139. — P. 104996.

12. Wu P., Song Y. Land use quantile regression modeling of fine particulate matter in Australia // *Remote Sensing*. — 2022. — Vol. 14, no. 6. — P. 1370.

13. Traffic noise modelling using land use regression model based on machine learning, statistical regression and GIS / Adulaimi A. A. A., Pradhan B., Chakraborty S., and Alamri A. // *Energies*. — 2021. — Vol. 14, no. 16. — P. 5095.

14. Prediction and evaluation of spatial distributions of ozone and urban heat island using a machine learning modified land use regression method / Han L., Zhao J., Gao Y., and Gu Z. // *Sustainable Cities and Society*. — 2022. — Vol. 78. — P. 103643.

15. Champendal A., Kanevski M., Huguenot P.-E. Air Pollution Mapping Using Nonlinear Land Use Regression Models // International Conference on Computational Science and Its Applications (ICCSA 2014). — 2014. — P. 682–690.

16. A land use regression model using machine learning and locally developed low-cost particulate matter sensors in Uganda / Coker E. S., Amegah A. K., Mwebaze E., Ssematimba J., and Bainomugisha E. // *Environmental Research*. — 2021. — Vol. 199. — P. 111352.

17. Using a land use regression model with machine learning to estimate ground level PM2.5 / Wong P. Y., Lee H. Y., Chen Y. C., Zeng Y. T., Chern Y. R., Chen N. T., Candice Lung S. C., Su H. J., and Wu C. D. // *Environmental Pollution*. — 2021. — Vol. 277. — P. 116846.

18. A comprehensive review of the development of land use regression approaches for modeling spatiotemporal variations of ambient air pollution: A perspective from 2011 to 2023 / Ma X., Zou B., Deng J., Gao J., Longley I., Xiao S., ... and Salmond, J. // *Environment International*. — 2024. — Vol. 183. — P. 108430.

19. von Klot S. Equivalence of using nested buffers and concentric adjacent rings as predictors in land use regression models // *Atmospheric Environment*. — 2011. — Vol. 45, no. 24. — P. 4108–4110.

20. Seinfeld J. H., Pandis S. N. Atmospheric chemistry and physics: from air pollution to climate change. — Hoboken : John Wiley & Sons, 2016.

21. Бызова Н. Л., Гаргер Е. К., Иванов В. Н. Экспериментальные исследования атмосферной диффузии и расчет распространения примеси. — Л. : Гидрометеоиздат, 1991.

22. Turner D. B. Workbook of atmospheric dispersion estimates: an introduction to dispersion modeling. — Boca Raton : CRC Press, 2020.

23. Danilov V. G., Maslov V. P., Volosov K. A. Mathematical modelling of heat and mass transfer processes. — Dordrecht : Springer, 2012. — Vol. 348.

24. Методика расчета концентраций в атмосферном воздухе вредных веществ, содержащихся в выбросах предприятий (ОНД-86): утв. Госкомгидрометом СССР 04.08.1986 № 192.

25. Air-Pollution modelling aspects: An overview / Chakraborty M., Bansal S., Masiwal R., Awasthi A. // *Air Pollution: Sources, Impacts and Controls*. — 2019. — P. 79–95.

26. Evaluation of roadway Gaussian plume models with large-scale measurement campaigns / Briant R., Seigneur C., Gadrat M., and Bugajny C. // *Geoscientific Model Development*. — 2013. — Vol. 6. — P. 445–456.

27. Zhang X., Wang J. Atmospheric dispersion of chemical, biological, and radiological hazardous pollutants: informing risk assessment for public safety // *Journal of Safety Science and Resilience*. — 2022. — Vol. 3, no. 4. — P. 372–397.

28. Comparative study of CALPUFF and CFD modeling of toxic gas dispersion in mountainous environments / Li M., Lo C., Yang D., Li Y., and Li Z. // *Atmosphere*. — 2024. — Vol. 15, no. 11. — P. 1370.

29. Матерон Ж. Основы прикладной геостатистики. — М. : Мир, 1968. — 407 с.

30. Burrough P. A., McDonnell R. A. Principles of geographical information systems. — Oxford : Oxford University Press, 1998. — 333 p.

31. Wackernagel H. Multivariate geostatistics: an introduction with applications. — Berlin : Springer, 2003. — 387 p.

32. Spatial interpolation of regional PM2.5 concentrations in China during COVID-19 incorporating multivariate data / Wei P., Xie S., Huang L., Liu L., Cui L., Tang Y., Meng C., and Zhang L. // *Atmospheric Pollution Research*. — 2023. — Vol. 14, no. 3. — P. 101688.

33. Using kriging incorporated with wind direction to investigate ground-level PM2.5 concentration / Zhang H., Zhan Y., Li J., Chao C. Y., Liu Q., Wang C., ... and Biswas P. // *Science of the Total Environment*. — 2021. — Vol. 751. — P. 141813.

34. Крюкова С. В., Симакина Т. Е. Пространственная интерполяция концентрации взвешенных частиц PM10 методом кокригинга // *Ученые записки Российского государственного гидрометеорологического университета*. — 2018. — № 51. — С. 150–161.

35. Diem J. E., Comrie A. C. Predictive mapping of air pollution involving sparse spatial observations // *Environmental Pollution*. — 2002. — Vol. 119, no. 1. — P. 99–117.

36. Birinci E., Deniz A., Özdemir E. T. The relationship between PM10 and meteorological variables in the mega city Istanbul // *Environmental Monitoring and Assessment*. — 2023. — Vol. 195, no. 2. — P. 304.

37. Breiman L. Random forests // *Machine Learning*. — 2001. — Vol. 45, no. 1. — P. 5–32.

38. Stekhoven D. J., Bühlmann P. MissForest — non-parametric missing value imputation for mixed-type data // *Bioinformatics*. — 2012. — Vol. 28, no. 1. — P. 112–118.

39. Park S. K. Assessing the impact of ozone and particulate matter on mortality rate from respiratory disease in Seoul, Korea // *Atmosphere*. — 2019. — Vol. 10, no. 11. — P. 685.

40. Predicting daily urban fine particulate matter concentrations using a random forest model / Brokamp C., Jandarov R., Hossain M., and Ryan P. // *Environmental Science & Technology*. — 2018. — Vol. 52, no. 7. — P. 4173–4179.

41. A multi-city air pollution population exposure study: combined use of chemical-transport and random forest models with dynamic population data / Gariazzo C., Carlino G., Silibello C. et al. // *Science of the Total Environment*. — 2020. — Vol. 724. — P. 138102.

42. Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain) / Sánchez A. S., Nieto P. G., Fernández P. R., del Coz Díaz J. J., and Iglesias-Rodríguez F. J. // *Mathematical and Computer Modelling*. — 2011. — Vol. 54, no. 5–6. — P. 1453–1466.

43. A systematic review of data mining and machine learning for air pollution epidemiology / Bellinger C., Mohomed Jabbar M. S., Zaïane O., and Osornio-Vargas A. // *BMC Public Health*. — 2017. — Vol. 17, no. 1. — P. 907.

44. PM2.5 estimation and spatial-temporal pattern analysis based on the modified support vector regression model and the 1 km resolution MAIAC AOD in Hubei, China / Chen N., Yang M., Du W., and Huang M. // *ISPRS International Journal of Geo-Information*. — 2021. — Vol. 10, no. 1. — P. 31.

45. Hornik K., Stinchcombe M., White H. Multilayer feedforward networks are universal approximators // *Neural Networks*. — 1989. — Vol. 2, no. 5. — P. 359–366.

46. Goodfellow I., Bengio Y., Courville A. Deep learning. — Cambridge : MIT Press, 2016.

47. Rumelhart D. E., Hinton G. E., Williams R. J. Learning representations by back-propagating errors // *Nature*. — 1986. — Vol. 323, no. 6088. — P. 533–536.

48. LeCun Y., Bengio Y., Hinton G. Deep learning // *Nature*. — 2015. — Vol. 521, no. 7553. — P. 436–444.

49. Explainable AI: interpreting, explaining and visualizing deep learning / Samek W., Montavon G., Vedaldi A., Hansen L. K., and Müller K.-R. — Cham : Springer, 2021.

50. A comparison of machine learning algorithms for mapping soil iron parameters indicative of pedogenic processes by hyperspectral imaging of intact soil profiles / Xu S., Zhao Y., Wang M., and Shi X. // *European Journal of Soil Science*. — 2022. — Vol. 73, no. 1. — P. e13204.

51. The evaluation on artificial neural networks (ANN) and multiple linear regressions (MLR) models for predicting SO2 concentration / Shams S. R., Jahani A., Kalantary S., Moeinaddini M., and Khorasani, N. // *Urban Climate*. — 2021. — Vol. 37. — P. 100837.

52. Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau / Dai F., Zhou O., Lv Z., Wang X., and Liu G. // *Ecological Indicators*. — 2014. — Vol. 45. — P. 184–194.

53. Spatial distribution and pollution assessment of heavy metals in urban soils from southwest China / Guo G., Wu F., Xie F., and Zhang R. // *Journal of Environmental Sciences*. — 2012. — Vol. 24, no. 3. — P. 410–418.

54. Digital soil mapping using artificial neural networks / Behrens T., Förster H., Scholten T., Steinrücken U., Spies E. D., and Goldschmitt M. // *Journal of Plant Nutrition and Soil Science*. — 2005. — Vol. 168, no. 1. — P. 21–33.

55. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping / Heung B., Ho H. C., Zhang J., Knudby A., Bulmer C. E., and Schmidt M. G. // Geoderma. — 2016. — Vol. 265. — P. 62–77.

56. Backpropagation applied to handwritten zip code recognition / LeCun Y., Boser B., Denker J. S., Henderson D., Howard R. E., Hubbard W., and Jackel L. D. // *Neural Computation*. — 1989. — Vol. 1, no. 4. — P. 541–551.

57. Levels, inventory, and risk assessment of heavy metals in wetland ecosystem, Northeast China: implications for snow cover monitoring / Zhang F., Meng B., Gao S., Hough R., Hu P., Zhang Z., ... and Cui S. // *Water*. — 2021. — Vol. 13, no. 16. — P. 2161.

58. Development of an LSTM broadcasting deep-learning framework for regional air pollution forecast improvement / Sun H., Fung J. C., Chen Y., Li Z., Yuan D., Chen W., and Lu X. // *Geoscientific Model Development*. — 2022. — Vol. 15, no. 22. — P. 8439–8452.

59. Антропов К.М., Вараксин А.Н. Оценка загрязнения атмосферного воздуха г. Екатеринбурга диоксидом азота методом Land Use Regression // *Экологические системы и приборы*. — 2011. — № 8. — С. 47–54.

60. Анализ статистических зависимостей распределения загрязняющих веществ в поверхностном слое почвы урбанизированных территорий с применением математических моделей (LUR метод) / Буевич А. Г., Сафина А. М., Сергеев А. П. и др. // *Геоэкология, инженерная геология, гидрогеология, геокриология*. — 2015. — № 3. — С. 268–279.

61. Градостроительный кодекс Российской Федерации : федеральный закон от 29.12.2004 № 190-ФЗ. — Ст. 30 «Правила землепользования и застройки». — URL: https://www.consultant.ru/document/cons_doc_LAW_51040/ (дата обращения: 23.04.2026).

62. О введении в действие Земельного кодекса Российской Федерации : федеральный закон от 08.08.2024 № 307-ФЗ. — URL: https://www.consultant.ru/document/cons_doc_LAW_480704/ (дата обращения: 23.04.2026).

63. ГОСТ Р 59055-2020. Охрана окружающей среды. Земли. Термины и определения. — URL: https://protect.gost.ru/v.aspx?control=8&id=205053 (дата обращения: 23.04.2026).

64. Report on the environment — land use / U.S. Environmental Protection Agency (EPA). — URL: https://www.epa.gov/report-environment/land-use (дата обращения: 23.04.2026).

65. Land use and land cover definitions / U.S. Geological Survey (USGS). — USGS Professional Paper 964, Appendix C. — URL: https://pubs.usgs.gov/publication/pp964 (дата обращения: 23.04.2026).

66. Glossary of land use and planning terms. — Barre City (USA). — URL: https://www.barrecity.org/client_media/files/PPA/Barre%20City%20Land%20Use%20Glossary%20Updated%2010-12-23.pdf (дата обращения: 23.04.2026).

67. Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems / Brauer M., Hoek G., van Vliet P., Meliefste K., Fischer P., Gehring U., ... and Brunekreef B. // *Epidemiology*. — 2003. — Vol. 14. — P. 228–239.

68. Establishing an air pollution monitoring network for intra-urban population exposure assessment: a location-allocation approach / Kanaroglou P. S., Jerrett M., Morrison J., Beckerman, B., Arain M. A., Gilbert N. L., and Brook J. R // *Atmospheric Environment*. — 2005. — Vol. 39, no. 13.

69. A land use regression model for predicting ambient fine particulate matter across Los Angeles, CA / Moore D. K., Jerrett M., Mack W. J., and Künzli N. // *Journal of Environmental Monitoring*. — 2007. — Vol. 9. — P. 246–252.

70. Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter / Henderson S., Beckerman B., Jerrett M., Brauer M. // *Environmental Science & Technology*. — 2007. — Vol. 41, no. 7. — P. 2422–2428.

71. Detecting urban land-use configuration effects on NO_2 and NO variations using geographically weighted land use regression / Song W., Jia H., Li Z., Tang D., and Wang C. // *Atmospheric Environment*. — 2019. — Vol. 197. — P. 166–176.

72. Comparison of land-use regression models between Great Britain and the Netherlands / Vienneau D., de Hoogh K., Beelen R., Fischer P., Hoek G., and Briggs D. // *Atmospheric Environment*. — 2010. — Vol. 44, no. 5. — P. 688–696.

73. Land use regression models for ultrafine particles in six European areas / van Nunen E., Vermeulen R., Tsai M. Y., Probst-Hensch N., Ineichen A., Davey M., … and Hoek G. // *Environmental Science & Technology*. — 2017. — Vol. 51, no. 6. — P. 3336–3345.

74. Development of NO_2 and PM land use regression models for estimating air pollution exposure in 36 study areas in Europe / Beelen R., Hoek G., Vienneau D., Eeftens

M., Dimakopoulou K., Pedeli X., ... and de Hoogh K. // *Atmospheric Environment*. — 2013. — Vol. 72. — P. 10–23.

75. Sheather S. A modern approach to regression with R. — New York : Springer, 2009. — 393 p.

76. Estimation of outdoor NOx, NO_2 and BTEX exposure in a cohort of pregnant women using land use regression modeling / Aguilera I., Sunyer J., Fernandez-Patier R., Hoek G., Aguirre-Alfaro A., Meliefste K., ... and Brunekreef B. // *Environmental Science & Technology*. — 2008. — Vol. 42. — P. 815–821.

77. A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments / Briggs D. J., de Hoogh C., Gulliver J., Wills J., Elliott P., Kingham S., and Smallbone K. // *Science of the Total Environment*. — 2000. — Vol. 253, no. 1–3. — P. 151–167.

78. Use of GIS and ancillary variables to predict volatile organic compound and nitrogen dioxide levels at unmonitored locations / Smith L., Mukerjee S., Gonzales M., Stallings C., Neas L., Norris G., and Özkaynak, H // *Atmospheric Environment*. — 2006. — Vol. 40. — P. 3773–3787.

79. Modeling annual benzene, toluene, NO_2, and soot concentrations on the basis of road traffic characteristics / Carr D., von Ehrenstein O., Weiland S., Wagner C., Wellie O., Nicolai T., and von Mutius E. // *Environmental Research*. — 2002. — Vol. 90. — P. 111–118.

80. Сергеев А. П., Баглаева Е. М., Субботина И. Е. Загрязнение почв города Тарко-Сале тяжелыми металлами // *Геоэкология. Инженерная геология. Гидрогеология. Геокриология*. — 2014. — № 1. — С. 28–36.

81. Ecological zoning of the Baikal Basin based on the results of chemical analysis of the composition of atmospheric precipitation accumulated in the snow cover / Molozhnikova Y. V., Shikhovtsev M. Y., Netsvetaeva O. G., and Khodzher T. V. // *Applied Sciences*. — 2023. — Vol. 13, no. 14. — Article 8171.

82. Kriging-based land-use regression models that use machine learning algorithms to estimate the monthly BTEX concentration / Hsu C. Y., Zeng Y. T., Chen Y. C., Chen

M. J., Lung S. C. C., and Wu C. D. // *International Journal of Environmental Research and Public Health*. — 2020. — Vol. 17, no. 19. — P. 6956.

83. Ren X., Mi Z., Georgopoulos P. G. Comparison of machine learning and land use regression for fine scale spatiotemporal estimation of ambient air pollution: modeling ozone concentrations across the contiguous United States // *Environment International*. — 2020. — Vol. 142. — P. 105827.

84. Bitta J., Svozilík V., Svozilikova Krakovska A. The neural network assisted land use regression // *Atmosphere*. — 2021. — Vol. 12, no. 4. — P. 452.

85. Jain S., Presto A. A., Zimmerman N. Spatial modeling of daily PM2.5, NO_2, and CO concentrations measured by a low-cost sensor network: comparison of linear, machine learning, and hybrid land use models // *Environmental Science & Technology*. — 2021. — Vol. 55, no. 13. — P. 8631–8641.

86. A land use regression model for ambient ultrafine particles in Montreal, Canada: a comparison of linear regression and a machine learning approach / Weichenthal S., Van Ryswyk K., Goldstein A., Bagg S., Shekkarizfard M., and Hatzopoulou M. // *Environmental Research*. — 2016. — Vol. 146. — P. 65–72.

87. Low-cost sensor networks and land-use regression: interpolating nitrogen dioxide concentration at high temporal and spatial resolution in Southern California / Weissert L., Alberti K., Miles E., Miskell G., Feenstra B., Henshaw G. S., ... and Williams, D. E. // *Atmospheric Environment*. — 2020. — Vol. 223. — P. 117287.

88. MapLUR: exploring a new paradigm for estimating air pollution using deep learning on map images / Steininger M., Kobs K., Zehe A., Lautenschlager F., Becker M., and Hotho A. // *ACM Transactions on Spatial Algorithms and Systems*. — 2020. — Vol. 6, no. 3. — P. 1–24.

89. Ripley B. D. Pattern recognition and neural networks. — Cambridge : Cambridge University Press, 1996.

90. Machine learning glossary / Google Developers // Google Machine Learning Guides. — 2023. — URL: https://developers.google.com/machine-learning/glossary (дата обращения: 24.04.2026).

91. Deep learning. — Coursera. — URL: https://www.coursera.org/specializations/deep-learning (дата обращения: 24.04.2026).

92. Data set terminology of deep learning in medicine: a historical review and recommendation / Walston S. L., Seki H., Takita H., Mitsuyama Y., Sato S., Hagiwara A., ... and Ueda D. // *Japanese Journal of Radiology*. — 2024. — Vol. 42. — P. 1100–1109.

93. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs / Gulshan V., Peng L., Coram M., Stumpe M. C., Wu D., Narayanaswamy A., ... and Webster D. R. // *The Journal of the American Medical Association*. — 2016. — Vol. 316. — P. 2402–2410.

94. Batchu V., Nearing G., Gulshan V. A deep learning data fusion model using Sentinel-1/2, SoilGrids, SMAP, and GLDAS for soil moisture retrieval // *Journal of Hydrometeorology*. — 2023. — Vol. 24. — P. 1789–1823.

95. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study / Chilamkurthy S., Ghosh R., Tanamala S., Biviji M., Campeau N. G., Venugopal V. K., ... and Warier P. // *The Lancet*. — 2018. — Vol. 392, no. 10162. — P. 2388–2396.

96. Artificial intelligence to detect papilledema from ocular fundus photographs / Milea D., Najjar R. P., Jiang Z., Ting D., Vasseneix C., Xu X., ... and Biousse, V. // *New England Journal of Medicine*. — 2020. — Vol. 382, no. 18. — P. 1687–1695.

97. Chest radiography as a biomarker of ageing: artificial intelligence-based, multi-institutional model development and validation in Japan / Mitsuyama Y., Matsumoto T., Tatekawa H., Walston S. L., Kimura T., Yamamoto A., ... and Ueda D. // *The Lancet Healthy Longevity*. — 2023. — Vol. 4, no. 9. — P. e478–e486.

98. ALT submission for OSACT shared task on offensive language detection / Hassan S., Samih Y., Mubarak H., Abdelali A., Rashed A., and Chowdhury S. A. // Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools. — 2020. — P. 61–65.

99. Applying deep learning in recognizing the femoral nerve block region on ultrasound images / Huang C., Zhou Y., Tan W., Qiu Z., Zhou H., Song Y., ... and Gao S. // *Annals of Translational Medicine*. — 2019. — Vol. 7, no. 18. — P. 453.

100. Development and validation of an artificial intelligence system for grading colposcopic impressions and guiding biopsies / Xue P., Tang C., Li Q., Li Y., Shen Y., Zhao Y., ... and Zhao F. // *BMC Medicine*. — 2020. — Vol. 18, no. 1. — P. 406.

101. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure / Roberts D. R., Bahn V., Ciuti S., Boyce M. S., Elith J., Guillera-Arroita G., ... and Dormann C. F. // *Ecography*. — 2017. — Vol. 40, no. 8. — P. 913–929.

102. Spatial validation reveals poor predictive performance of large-scale ecological mapping models / Ploton P., Mortier F., Réjou-Méchain M., Barbier N., Picard N., Rossi V., ... and Pélissier R. // *Nature Communications*. — 2020. — Vol. 11. — P. 4540.

103. A new technique for evaluating land-use regression models and their impact on health effect estimates / Wang M., Brunekreef B., Gehring U., Szpiro A., Hoek G., and Beelen R. // *Epidemiology*. — 2016. — Vol. 27, no. 1. — P. 51–56.

104. blockCV: an R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models / Valavi R., Elith J., Lahoz-Monfort J. J., and Guillera-Arroita G. // *Methods in Ecology and Evolution*. — 2019. — Vol. 10. — P. 225–232.

105. Wang Y., Khodadadzadeh M., Zurita-Milla R. Spatial+: a new cross-validation method to evaluate geospatial machine learning models // *International Journal of Applied Earth Observation and Geoinformation*. — 2023. — Vol. 121. — P. 103364.

106. Abbasian M., Moghim S., Abrishamchi A. Performance of the general circulation models in simulating temperature and precipitation over Iran // *Theoretical and Applied Climatology*. — 2019. — Vol. 135. — P. 1465–1483.

107. Modelling monthly rainfall of India through transformer-based deep learning architecture / Nayak G. H. H., Alam W., Singh K. N., Avinash G., Ray M., and Kumar R. R. // *Modeling Earth Systems and Environment*. — 2024. — Vol. 10, no. 3. — P. 3119–3136.

108. Modelling and prediction of major soil chemical properties with Random Forest: machine learning as tool to understand soil-environment relationships in Antarctica / Siqueira R. G., Moquedace C. M., Fernandes-Filho E. I., Schaefer C. E., Francelino M. R., Sacramento I. F., and Michel R. F. // *Catena*. — 2024. — Vol. 235. — P. 107677.

109. Taylor A. B., MacKinnon D. P. Four applications of permutation methods to testing a single-mediator model // *Behavior Research Methods*. — 2012. — Vol. 44, no. 3. — P. 806–844.

110. Taylor K. E. Summarizing multiple aspects of model performance in a single diagram // *Journal of Geophysical Research: Atmospheres*. — 2001. — Vol. 106, no. D7. — P. 7183–7192.

111. Wavelet neural networks and gene expression programming models to predict short-term soil temperature at different depths / Samadianfard S., Asadi E., Jarhan S., Kazemi H., Kheshtgar S., Kisi O., ... and Manaf A. A. // *Soil and Tillage Research*. — 2018. — Vol. 175. — P. 37–50.

112. Application of Taylor diagram in the evaluation of joint environmental distributions' performances / Simão M. L., Videiro P. M., Silva P. B. A., de Freitas Assad L. P., and Sagrilo L. V. S // *Marine Systems & Ocean Technology*. — 2020. — Vol. 15. — P. 151–159.

113. Good P. Permutation tests: a practical guide to resampling methods for testing hypotheses. — New York : Springer, 2013.

114. Berry K. J., Johnston J. E., Mielke P. W. Jr. Permutation methods // Wiley Interdisciplinary Reviews: Computational Statistics. — 2011. — Vol. 3, no. 6. — P. 527–542.

115. Efron B., Tibshirani R. J. An introduction to the bootstrap. — New York : Chapman & Hall, 1993.

116. Manly B. F. J. Randomization, bootstrap and Monte Carlo methods in biology. — Boca Raton : Chapman & Hall/CRC, 2007.

117. Fisher R. A. The design of experiments. — Edinburgh : Oliver and Boyd, 1935.

118. Ross S. M. Simulation, bootstrap statistical methods, and permutation tests // Introduction to probability and statistics for engineers and scientists. — Amsterdam : Academic Press, 2021. — P. 619–647.

119. Fişek M. H., Barlas Z. Permutation tests for goodness-of-fit testing of mathematical models to experimental data // *Social Science Research*. — 2013. — Vol. 42, no. 2. — P. 482–495.

120. Quenouille M. H. Approximate tests of correlation in time-series // *Journal of the Royal Statistical Society. Series B (Methodological)*. — 1949. — Vol. 11, no. 1. — P. 68–84.

121. Tukey J. W. Bias and confidence in not quite large samples // *Annals of Mathematical Statistics*. — 1958. — Vol. 29, no. 2. — P. 614–623.

122. Hastie, T., Tibshirani, R., Friedman, J. The elements of statistical learning. — 2009.

123. Golub G. H., Van Loan C. F. Matrix computations. — Baltimore : Johns Hopkins University Press, 2013.

124. Ramsey C. A., Hewitt A. D. A methodology for assessing sample representativeness // *Environmental Forensics*. — 2005. — Vol. 6, no. 1. — P. 71–75.

125. Peterson S. A., Urquhart N. S., Welch E. B. Sample representativeness: a must for reliable regional lake condition estimates // *Environmental Science & Technology*. — 1999. — Vol. 33, no. 10. — P. 1559–1565.

126. Choice and criteria for selection of sampling strategies in environmental radioactivity monitoring / Scott E. M., Dixon P., Voigt G., and Whicker W. // *Applied Radiation and Isotopes*. — 2008. — Vol. 66, no. 11. — P. 1575–1581.

127. Representativeness-based sampling network design for the State of Alaska / Hoffman F. M., Kumar J., Mills R. T., and Hargrove W. W. // *Landscape Ecology*. — 2013. — Vol. 28, no. 8. — P. 1567–1586.

128. Ghorbani, A., Zou, J. Data shapley: Equitable valuation of data for machine learning // *International conference on machine learning.* — 2019. — P. 2242–2251.

129. Kwon Y., Zou J. Data-OOB: out-of-bag estimate as a simple and efficient data value // *International Conference on Machine Learning*. — 2023. — P. 18135–18152.

130. Zhang J., Zhang C. Sampling and sampling strategies for environmental analysis // *International Journal of Environmental Analytical Chemistry*. — 2012. — Vol. 92, no. 4. — P. 466–478.

131. Counter-prediction approach to predict the missing values of a spatial series on the example of the dustiness in the snow cover / Sergeev A. P., Shichkin A. V., Buevich A. G., and Baglaeva E. M. // Modeling Earth Systems and Environment. — 2023. — Vol. 9. — No. 2. – P. 1523–1530.

132. Пылегрязевые отложения как индикатор эколого-геохимического состояния территории арктического города / Ярмошенко И.В., Шевченко А.В., Глухов В.С., Малиновский Г.П., Селезнев А.А. // Арктика: экология и экономика. — 2024. — Т. 14, — № 3. — С. 449–461.

133. The effect of splitting of raw data into training and test subsets on the accuracy of predicting spatial distribution by a multilayer perceptron / Baglaeva E. M., Sergeev A. P., Shichkin A. V., and Buevich A. G. // Mathematical Geosciences. — 2020. — Vol. 52. — No. 1. — P. 111-121.

Список работ, опубликованных по теме диссертации

Статьи в журналах перечня ВАК и в изданиях, индексируемых Scopus и Web of Science

1. Sergeev A. P., Shichkin A. V., Buevich A. G., **Butorova A. S.**, Baglaeva E. M. Using land use methodology to construct ring spatial variables for modeling and mapping spatial distribution of dust in snow cover // The European Physical Journal Special Topics. — 2025. — Vol. 234. — P. 4607–4618. — DOI: 10.1140/epjs/s11734-024-01341-w (WoS, Scopus Q2)

2. Sergeev A. P., Shichkin A. V., Baglaeva E. M., Buevich A. G., **Butorova A. S.** A permutation approach to evaluating the performance of a forecasting model of methane content in the atmospheric surface layer of arctic region // Atmospheric Pollution Research. — 2024. — Vol. 15, Issue 2. — Article No. 102000. — DOI: 10.1016/j.apr.2023.102000 (WoS, Scopus Q2)

3. Sergeev A. P., **Butorova A. S.**, Shichkin A. V., Buevich A. G., Baglaeva E. M. Increasing the informativeness of performance assessment of predictive models of heavy metal spatial distributions in the topsoil by permutation approach // Modeling Earth System and Environment. — 2024. — Vol. 10. — P. 4387–4400. — DOI: 10.1007/s40808-024-02034-y (WoS, Scopus Q2)

4. **Буторова А. С.** Применение LUR – подхода к моделированию поверхностного распределения пыли в снеговом покрове южной части г. Новый Уренгой / А. С. Буторова, А. В. Шичкин, А. П. Сергеев, А. Г. Буевич, Е. М. Баглаева // Теоретическая и прикладная экология. — 2025. — № 4. — С. 37–43. — DOI: 10.25750/1995-4301-2025-4-037-043 (Scopus Q3, WoS)

5. **Butorova A. S.**, Sergeev A. P. Land Use Segmentation Approach for Spatial Modeling in Complex Heterogeneous Environments // The European Physical Journal Special Topics — 2026. — DOI: https://doi.org/10.1140/epjs/s11734-026-02474-w (WoS, Scopus Q2)

6. **Буторова А. С.**, Сергеев А. П. Численное исследование различий между асимптотическим и пермутационным подходами к оценке качества модели при

малом числе наблюдений // Южно-Сибирский научный вестник. — 2026. — № 2. — С. 38–44.

Публикации в сборниках трудов конференций, индексируемых Scopus и Web of Science

7. **Butorova A. S.**, Sergeev A. P. Land Use Approach for High-Resolution Spatial Modeling of Industrial Emissions // 2026 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT). — 2026. — P. 1–4. — DOI: 10.1109/USBEREIT70063.2026.11580492 (Scopus).

8. **Butorova A. S.**, Sergeev A. P. A Computational Framework for Land Use Modeling of the Spatial Distribution of Impurities in the Environment // 2025 9th Scientific School “Dynamics of Complex Networks and their Applications” (DCNA). Kaliningrad, Russian Federation, 2025. — P. 32–34. — DOI: 10.1109/DCNA67486.2025.11199925 (Scopus).

9. An Improved Land Use Methodology Based on Machine Learning for Predicting the Spatial Distribution of Contaminants in Urban Surface Sediments / **A. Butorova**, A. Buevich, A. Sergeev, A. Shichkin, E. Baglaeva, A. Seleznev, I. Subbotina // 2025 9th Scientific School “Dynamics of Complex Networks and their Applications” (DCNA). Kaliningrad, Russian Federation, September 2025. — P. 28–31. — DOI: 10.1109/DCNA67486.2025.11199956 (Scopus).

10. Evaluation of the models of copper spatial distribution in the surface layer of the soil based on artificial neural networks by the permutation method / A. P. Segeev, **A. S. Butorova**, A. V. Shichkin, E. M. Baglaeva, A. G. Buevich, I. E. Subbotina, M. V. Sergeeva // International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2022). Heraklion, Greece: AIP, 2024. — V. 3094. — Article number 190003. — DOI: 10.1063/5.0210522 (Scopus).

Патенты на изобретения

11. Сергеев А. П., **Буторова А. С.**, Корюкин Е.А. Способ построения полей пространственного распределения примесей — Патент на изобретение RU 2 841 093 от 02.06.2025. Бюл. № 16. Заявка № 2024122810 от 09.08.2024.

12. Сергеев А. П., **Буторова А. С.**, Буевич А. Г., Субботина И. Е. Способ подготовки пространственных переменных для построения модели распределения примесей в верхнем слое почвы — Патент на изобретение RU 2 863 777 от 09.06.2026. Бюл. № 16. Заявка № 2024121454/20(047654) от 26.07.2024.

Свидетельства о государственной регистрации программ для ЭВМ

13. **Буторова А. С.**, Сергеев А. П. Программный комплекс для Land Use моделирования — Свидетельство о гос. регистрации программы для ЭВМ № 2026666809; дата регистрации 03.06.2026.

Прочие публикации

14. Баглаева Е. М., Буевич А. Г., Шичкин А. В., Сергеев А. П., **Буторова А. С.** Гибридное моделирование на основе LUR-подхода вариаций распределения тяжёлых металлов в верхнем слое почвы на примере города Тарко-Сале // Геоэкология. Инженерная геология, гидрогеология, геокриология. — 2025. — № 1. — С. 87–96. — DOI: 10.31857/S0869780925010097

15. **Буторова А. С.**, Шичкин А. В., Сергеев А. П., Баглаева Е. М., Буевич А. Г. Моделирование пространственного распределения хрома и марганца в почве: подбор обучающего подмножества // Геоэкология. Инженерная геология, гидрогеология, геокриология. — 2023. — № 5. — С. 88–96. — DOI: 10.31857/S0869780923050028

16. **Буторова А. С.**, Сергеев А. П., Шичкин А. В., Буевич А. Г., Баглаева Е. М., Субботина И. Е. Применение перестановочного метода к оценке прогностической способности моделей пространственного распределения концентраций меди и железа в верхнем слое почвы // Геоинформатика. — 2022. — № 2. — С. 42–53. — DOI: 10.47148/1609-364X-2022-2-42-53

17. Баглаева Е. М., Сергеев А. П., Шичкин А. В., Буевич А. Г., Субботина И. Е., **Буторова А. С.** Сравнительный анализ алгоритмов разделения данных для обучения искусственных нейронных сетей при моделировании пространственного распределения тяжелых металлов в верхнем слое почвы // Траектория исследований – человек, природа, технологии. — 2022. — № 3 (3). — С. 73–88.

18. Сергеев А. П., **Буторова А. С.**, Корюкин Е. А., Бобаков В. С., Павлова С. В. Непараметрический рандомизационный тест: идея, алгоритм, преимущества, недостатки и пример применения // Траектория исследований – человек, природа, технологии. — 2023. — № 4 (8). — С. 54–67.

Приложение А

Подбор гиперпараметров для моделей Land Use Buffers и Land Use Rings

Таблица А.1 Сетки гиперпараметров моделей Land Use Buffers и Land Use Rings

Модель	Гиперпараметр	Значения
Стандартная линейная регрессия	–	–
Ridge	коэффициент регуляризации α	{0,01; 0,1; 1; 10; 100}
Lasso	коэффициент регуляризации α	{0,0001; 0,001; 0,01; 1; 10}
Random Forest (RF)	число деревьев *n_estimators*	{50; 100; 200}
	максимальная глубина деревьев *max_depth*	{None; 5; 10; 20}
	минимальное число точек в узле, при котором разбиение еще разрешено, *min_samples_split*	{2; 5; 10}
	минимальное число точек в листе – конечном узле дерева, в котором формируется прогноз, *min_samples_leaf*	{1; 2; 4}
Gradient Boosting (GB)	число деревьев *n_estimators*	{50; 100; 200}
	скорость обучения *learning_rate*	{0,01; 0,05; 0,1}
	максимальная глубина *max_depth*	{2; 3; 5}
	минимальное число точек для разбиения *min_samples_split*	{2; 5}
	минимальное число точек в листе *min_samples_leaf*	{1; 2}

Таблица А.1 (Продолжение)

Модель	Гиперпараметр	Значения
Support Vector Regression (SVR)	ядро *kernel*	{linear; rbf; poly}
	параметр регуляризации C	{0,1; 1; 10; 50; 100}
	ширина ε-трубки	{0,01; 0,1; 0,5; 1}
	параметр ядра γ	{'scale'; 'auto'}
k-Nearest Neighbors (KNN)	число соседей k	{2; 3; 5; 7; 10}
	веса	{uniform; distance}
	метрика Минковского p	{1; 2}, где $p = 1$ – манхэттенское расстояние, $p = 2$ – евклидово
Multilayer Perceptron (MLP)	размер скрытого слоя	{(5,); (10,); (15,)}
	функция активации	{relu; tanh}
	L2-регуляризация α	{0,0001; 0,001; 0,01}
	скорость обучения	{0,001; 0,01}
	оптимизатор	'lbfgs'

Приложение Б

Дополнительные метрики оценки качества моделей

Здесь O – наблюдаемый набор данных, P – предсказанный набор данных, n – число наблюдений.

Средний квадрат ошибки

$$MSE = \frac{\sum(P-O)^2}{n}$$

Средняя абсолютная ошибка в процентах

$$MAPE = \frac{1}{n}\sum\left|\frac{P-O}{O}\right| * 100\%$$

Среднеквадратическая относительная ошибка

$$RMSRE = \sqrt{\frac{1}{n}\sum\left(\frac{P-O}{O}\right)^2}$$

Относительная среднеквадратическая ошибка

$$RRMSE = \frac{\sqrt{\sum(P-O)^2}}{\sqrt{\sum O^2}}$$

Нормализованная среднеквадратическая ошибка

$$NRMSE = \frac{\sqrt{\sum(P-O)^2}}{\sqrt{\sum(O-\bar{O})^2}}$$

Стандартное отклонение

$$\sigma_O = \sqrt{\frac{\sum(O-\bar{O})^2}{n-1}}, \sigma_P = \sqrt{\frac{\sum(P-\bar{P})^2}{n-1}},$$

где σ_O – стандартное отклонение наблюдаемого набора данных, σ_P – стандартное отклонение предсказанного набора данных.

Индексы согласия Уиллмотта

$$IA1 = 1 - \frac{\sum|P-O|}{\sum(|P-\bar{O}|+|O-\bar{O}|)},$$

$$IA2 = 1 - \frac{\sum(P-O)^2}{\sum(|P-\bar{O}|+|O-\bar{O}|)^2},$$

Индексы $IA1$ и $IA2$ являются стандартизированными показателями степени ошибки прогнозирования модели и находятся в диапазоне от 0 до 1, где значение 1 указывает на полное совпадение, а 0 – на полное отсутствие согласия.

Приложение В

Дополнительные метрики качества моделей

Таблица В.1 Дополнительные метрики качества моделей на синтетических данных

Модель Land Use		*MSE*	*MAPE*	*RMSRE*	*RRMSE*	*NRMSE*	σ_O	σ_P	*IA1*	*IA2*
Linear	Buffers	4,132	43,049	1,758	0,094	0,182	11,140	11,491	0,915	0,992
	Rings	4,132	43,049	1,758	0,094	0,182	11,140	11,491	0,915	0,992
Ridge	Buffers	1,961	35,167	1,493	0,065	0,125	11,140	11,240	0,941	0,996
	Rings	2,574	20,357	0,666	0,074	0,143	11,140	10,974	0,931	0,995
Lasso	Buffers	3,320	41,602	1,731	0,084	0,163	11,140	11,220	0,923	0,993
	Rings	4,009	42,054	1,713	0,093	0,179	11,140	11,475	0,916	0,992
RF	Buffers	9,039	74,613	3,363	0,131	0,252	11,140	10,279	0,887	0,980
	Rings	18,020	95,329	4,168	0,192	0,368	11,140	9,277	0,823	0,956
GB	Buffers	9,264	84,724	3,931	0,137	0,265	11,140	10,844	0,881	0,981
	Rings	15,108	89,054	3,868	0,177	0,342	11,140	10,209	0,839	0,966
kNN	Buffers	11,174	78,499	3,620	0,145	0,277	11,140	10,171	0,879	0,976
	Rings	22,512	143,617	6,564	0,215	0,413	11,140	9,374	0,801	0,945
SVR	Buffers	2,998	32,928	1,395	0,077	0,150	11,140	11,323	0,931	0,994
	Rings	2,489	26,215	1,004	0,073	0,140	11,140	11,139	0,934	0,995
MLP	Buffers	30,055	142,439	6,451	0,250	0,485	11,140	12,869	0,801	0,946
	Rings	26,099	129,076	5,629	0,236	0,456	11,140	12,011	0,795	0,949
Land Use Segmentation		1,182	19,261	0,826	0,050	0,097	11,112	11,068	0,954	0,998

Таблица В.2 Дополнительные метрики качества моделей на данных наблюдений потока пыли в снеговом покрове г. Новый Уренгой

Модель Land Use		*MSE*	*MAPE*	*RMSRE*	*RRMSE*	*NRMSE*	σ_O	σ_P	*IA1*	*IA2*
Ridge	Buffers	0,001	94,612	2,084	0,379	0,732	0,036	0,022	0,525	0,784
	Rings	0,001	95,694	2,066	0,401	0,773	0,036	0,024	0,523	0,773
ElasticNet	Buffers	0,001	99,564	2,242	0,378	0,728	0,036	0,021	0,521	0,781
	Rings	0,001	99,641	2,237	0,378	0,73	0,036	0,021	0,522	0,782
RF	Buffers	0,001	81,983	1,589	0,457	0,881	0,036	0,028	0,536	0,738
	Rings	0,001	80,747	1,597	0,443	0,855	0,036	0,03	0,563	0,769
GB	Buffers	0,001	95,811	1,943	0,484	0,933	0,036	0,028	0,482	0,700
	Rings	0,001	95,616	1,907	0,511	0,986	0,036	0,031	0,497	0,691
kNN	Buffers	0,001	83,394	1,658	0,437	0,843	0,036	0,024	0,495	0,728
	Rings	0,001	90,952	1,825	0,490	0,944	0,036	0,028	0,521	0,700
Land Use Segmentation		0,000	38,755	0,659	0,273	0,526	0,036	0,031	0,748	0,920

Приложение Г

Примеры применения пермутационного подхода к моделям с различными типами связи между X и Y

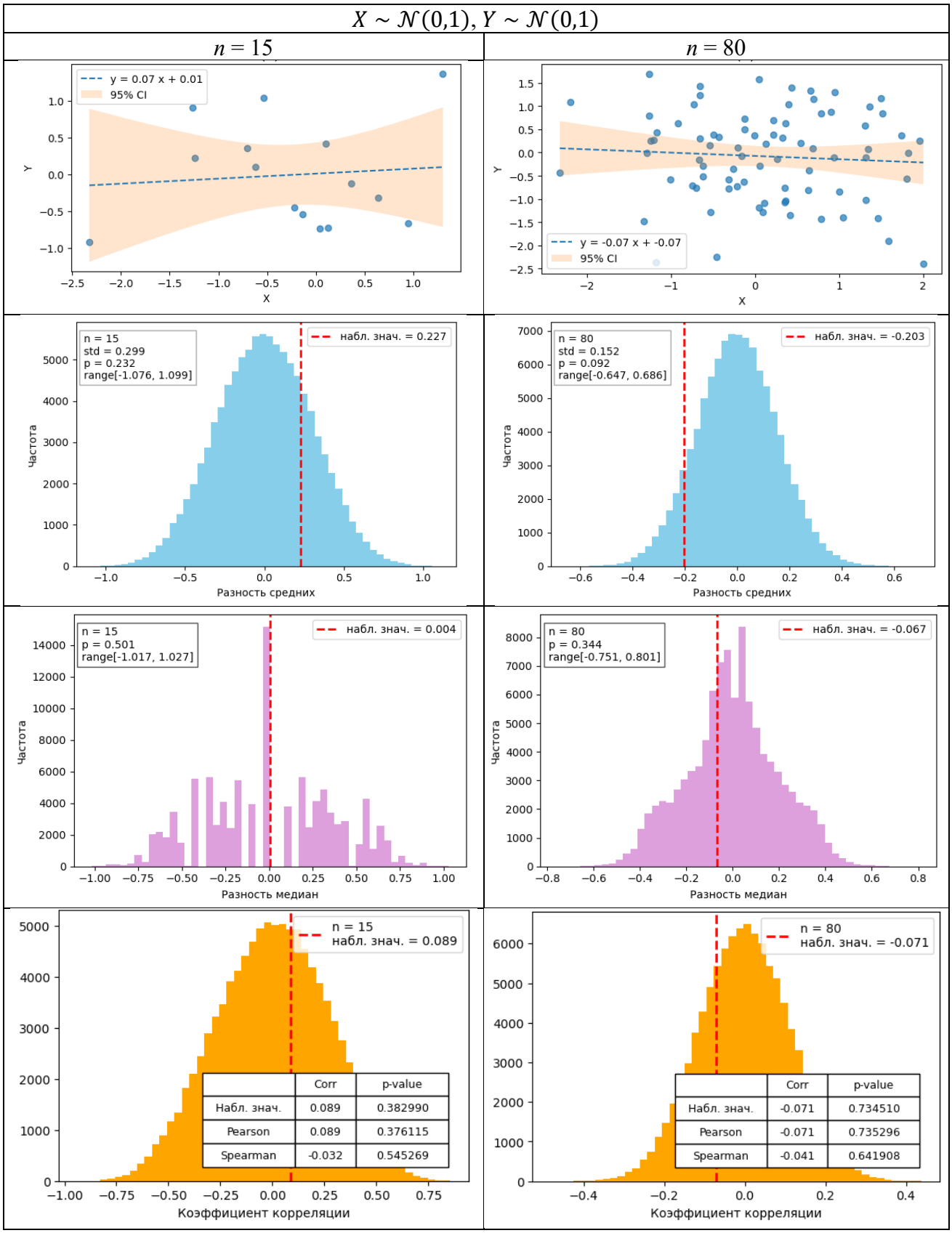


Рис. Г.1 Отсутствие связи между X и Y

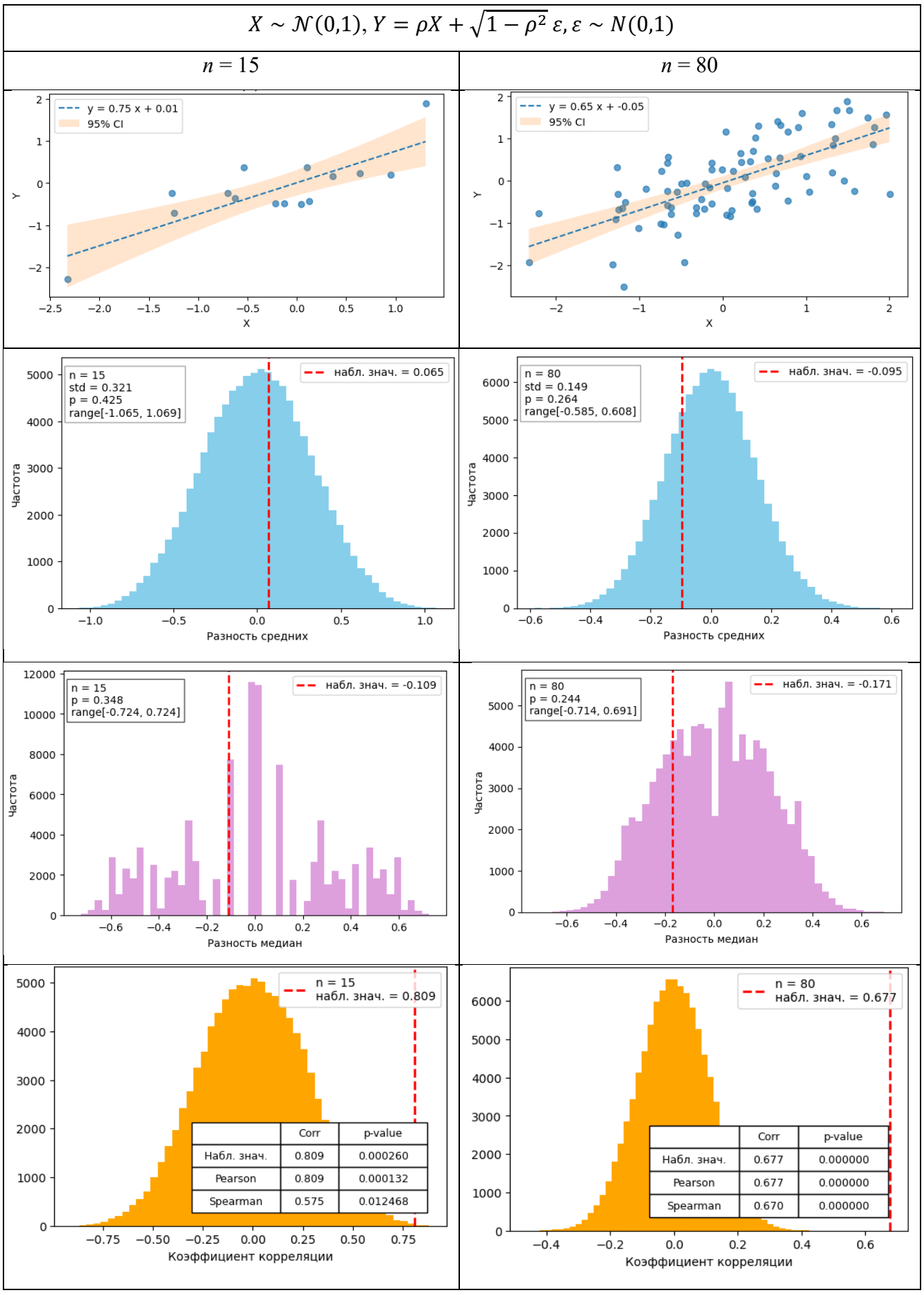


Рис. Г.2 Линейная связь между X и Y

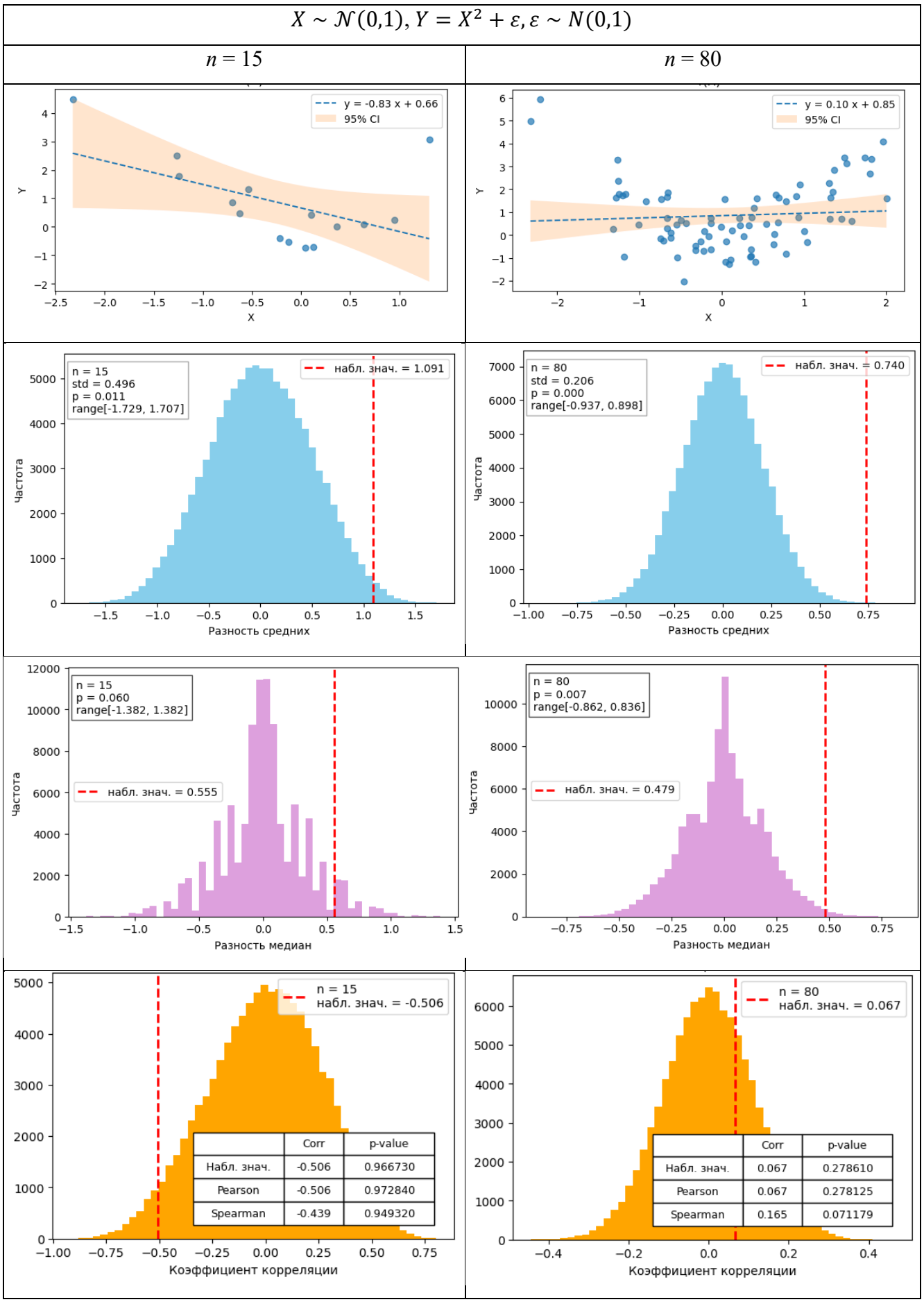


Рис. Г.3 Нелинейная (квадратичная) связь между X и Y

$$X = \begin{cases} 0 + \varepsilon_1, & i \le n/2 \\ 2 + \varepsilon_1, & i > n/2 \end{cases}, \varepsilon_1 \sim N(0, \sigma^2), Y = \rho X + \sqrt{1 - \rho^2}\, \varepsilon_2, \varepsilon_2 \sim N(0,1)$$

$n = 15$	$n = 80$

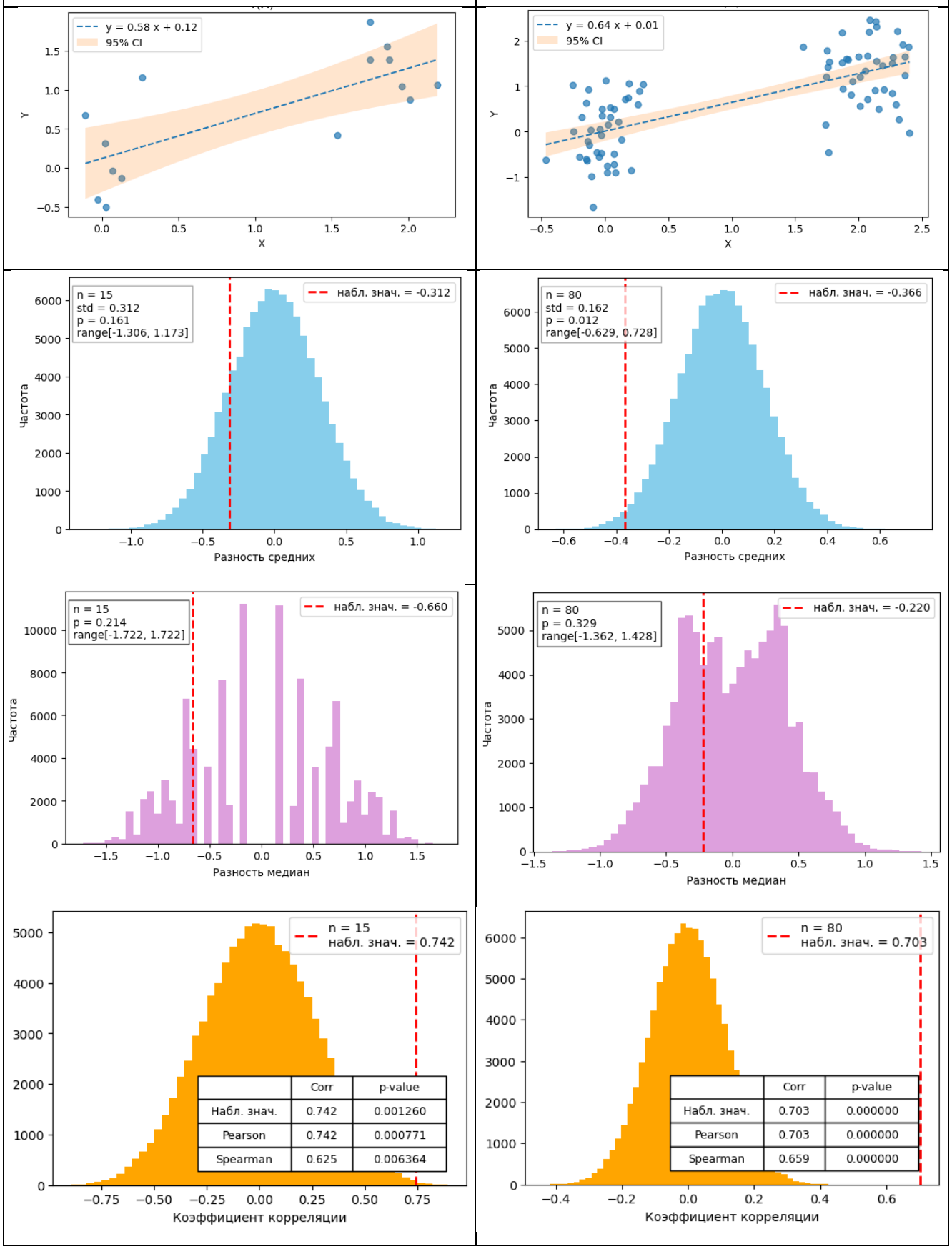


Рис. Г.4 Линейная связь между X и Y, значения X формируют кластеры