

## Опыт применения моделей машинного обучения для прогнозирования численности мелких грызунов (на примере *Myodes rutilus*)

Ямборко А.В.<sup>\*1,2,3</sup>, Тимошилов В.И.<sup>†3,4</sup>

<sup>1</sup>Федеральный центр анализа и оценки техногенного воздействия, Научно-исследовательский центр по редким и исчезающим видам животных и растений (филиал), Москва, Россия

<sup>2</sup>Институт биологических проблем Севера ДВО РАН, Магадан, Россия

<sup>3</sup>Московский физико-технический институт, Москва, Россия

<sup>4</sup>Курский государственный медицинский университет, Курск, Россия

**Аннотация.** Проведен сравнительный анализ регрессионных моделей машинного обучения для прогнозирования изменений относительной численности природных популяций красной полевки на один год вперед. Для обучения, валидации и тестирования моделей случайного леса и многослойного персептрона использовали демографические и климатические данные для восточного сектора Субарктики в целом, а также отдельно для двух пунктов региона. Каждый набор данных, содержащий ежегодные наблюдения за различными показателями, в модели представлен как вектор признаков (целевой признак – относительная численность на один год вперед; предикторы – популяционные и климатические данные), без привязки к временной структуре. При традиционном прогнозировании временных рядов будущие значения получают на основе последовательности прошлых значений, в то время как предсказание по векторам рассматривает каждое наблюдение как отдельную точку в пространстве признаков, формируя векторные представления данных. Показано, что многослойный персептрон демонстрирует лучшие результаты и точность прогноза по всем выборкам. Случайный лес характеризуется меньшей устойчивостью и точностью. В целом по региону и отдельным пунктам предпочтительнее использовать нейросетевые методы, например, многослойный персептрон. В случае необходимости выполнения быстрого моделирования или интерпретируемости можно использовать модель случайного леса. На примере популяций красной полевки показано, что точный прогноз относительной численности мелких грызунов на один год вперед можно получить, используя ограниченный по времени набор данных о состоянии популяций и условиях их обитания. Модели машинного обучения могут применяться для решения задач эпидемиологии и защиты растений.

**Ключевые слова:** машинное обучение, случайный лес, многослойный персептрон, прогноз численности, *Myodes rutilus*

### ВВЕДЕНИЕ

Представители подсемейств мышиных, полевковых и др. известны как вредители культурных растений и древесных пород, носители природно-очаговых зоонозов, чем создают угрозу для здоровья людей и домашних животных, наносят прямой урон сельскому и лесному хозяйству.

\* yambor84@inbox.ru

† timoshilovvi@kursksmu.net

С другой стороны, как консументы первого порядка, грызуны занимают важное место в переносе энергии по пищевым цепям в природных экосистемах, так как составляют основу кормовой базы мелких и средних хищных позвоночных животных. Высокая плодовитость, быстрое половое созревание, и, в то же время, короткая продолжительность жизни и высокая смертность мелких грызунов обуславливают значительные по амплитуде колебания численности от года к году: периоды низкого обилия сменяются подъемами («вспышками») численности. Это обуславливает необходимость разработки точных прогнозов обилия грызунов в целях управления рисками и минимизации последствий их негативного воздействия при высокой численности (потрава урожая культур и саженцев при лесовосстановлении, распространение инфекционных болезней и т.п.).

Красная полевка (*Myodes rutilus* (Pallas, 1779)) – широко распространённый и наиболее многочисленный в Северной Азии вид мелких грызунов, с одной стороны, важный элемент энергетической и трофической цепей в экосистемах, а с другой – вредитель лесного хозяйства и носитель природно-очаговых инфекций. Подъемы численности этого вида обычно происходят с периодичностью 3–5 лет.

Считается, что подъемы и спады численности грызунов обусловлены влиянием внешних факторов (абиотических и биотических условий) [1, 2, 3], а также эндогенной регуляцией численности популяций, когда плотность собственного населения становится регулирующим фактором (механизм исключает перенаселение и перерасход ресурсов) [4]. Показано, что динамика популяции зависит от начальных условий или текущих значений численности [5]. В настоящее время, популярным считается многофакторный подход к проблеме [6, 7], когда на динамику численности мелких грызунов воздействует множество факторов, как внутренних (плотность населения, возрастная структура и т.д.), так и внешних, включая погодные условия [8], доступность кормов и межпопуляционные взаимодействия (конкуренция, хищничество, инфекции и инвазии).

Прогнозирование в экологии включает в себя использование математических моделей, статистических методов и вычислительных инструментов для анализа данных и выявления закономерностей [9]. Машинное обучение – область искусственного интеллекта, которая фокусируется на разработке алгоритмов, позволяющих компьютерам обучаться на данных и делать прогнозы без явного программирования. Несмотря на успехи, достигнутые в других областях, при использовании искусственного интеллекта в экологии естествоиспытатели сталкиваются с проблемами, такими как нехватка данных или их низкое качество. Более того, нейронные сети и алгоритмы глубокого обучения являются в своем роде «черными ящиками», что затрудняет понимание их работы, а, следовательно, и интерпретацию результатов [10]. В контексте прогнозирования численности грызунов ключевое отличие нейросетевых моделей от традиционных подходов заключается в том, что первые позволяют использовать факторы непосредственно в качестве предикторов, а релевантность оценивается самими моделями [11].

Популярными математическими средствами изучения динамики биологических популяций является имитационное моделирование, а для прогнозирования применяют методы анализа временных рядов. Установлено, что алгоритмы машинного обучения показывают лучшие результаты, чем имитационные демографические модели [12], а анализ временных рядов можно проводить, непосредственно используя машинное обучение с запоминанием прошлых значений, например, рекуррентную нейронную сеть [13]. Синергия искусственного интеллекта и математических методов показала эффективность анализа демографических параметров для получения информации о размере, состоянии и поведении популяций [14]. В зоопаразитологических исследованиях алгоритмы машинного обучения показали высокую точность

прогнозирования по сравнению с традиционными методами анализа временных рядов [15].

Внутренняя структура экологических систем слишком сложна для полноценного описания имитационной моделью, а временные ряды работ по учету грызунов непродолжительны (обычно несколько десятков лет) и часто имеют пропуски в годах наблюдений. Всё это обуславливает необходимость применение алгоритмов, способных на ограниченных в количестве (короткие временные ряды) и не последовательных данных (пропуски в исследованиях) получить точный прогноз численности мелких грызунов в краткосрочной перспективе, например, на один год вперед, чего достаточно для принятия решений на практике.

В отечественной практике биоэкологического прогнозирования алгоритмы машинного обучения до сих пор не нашли широкого применения. Тем не менее, они были протестированы для прогнозирования активности природных очагов инфекций. На основе нейронных сетей построены модели прогнозирования эпизоотий туляремии [16], а также заболеваемости человека зоонозами – геморрагической лихорадкой с почечным синдромом, лептоспирозом, иксодовым клещевым боррелиозом [17, 18].

Цель настоящей работы – провести сравнительный анализ и оценить эффективность регрессионных моделей машинного обучения для прогнозирования изменений численности природной популяции красной полевки на один год вперед.

## МАТЕРИАЛЫ И МЕТОДЫ

*Зоологические данные.* Использованы материалы, собранные в разные периоды в восточном секторе Субарктики, хранящиеся в Институте биологических проблем ДВО РАН (г. Магадан) [19]. Наименование пунктов, координаты и годы исследований представлены в таблице 1.

**Таблица 1.** Пункты исследований популяций красной полевки

Пункт исследований	Координаты		Период исследований, годы / количество лет	Отловлено животных, штук
	с.ш.	в.д.		
Магадан	59°57'	150°94'	2011–2015 / 5	460
Челомджа	60°17'	147°37'	1980–1989 / 10	3204
Кулу	61°47'	146°00'	1981–1984 / 4	394
Буюнда	62°26'	153°20'	2001–2009 / 9	3481
Омолон	66°05'	159°10'	1972–1975 / 4	2688
Анадырь	64° 55'	168°34'	1986–1989 / 4	565

Учеты численности животных проводили методом безвозвратного изъятия с помощью ловушек Геро (25–50 штук), возраст определяли по состоянию корней зубов, состоянию репродуктивной системы и наличию вилочковой железы. Самки считались фертильными (половозрелыми) при наличии эструса, текущей (эмбрионы) и следов прошлой беременности (плацентарные пятна). Для построения моделей прогнозирования использовали три демографических параметра [7]:

1) относительная численность животных:

$$N = \left( \frac{n}{T \cdot D} \right) \cdot 100, \quad (1)$$

где  $n$  – количество добытых животных, штук;  $T$  – количество ловушек, штук;  $D$  – продолжительность наблюдения, суток;

2) доля участвующих в размножении самок-сеголеток:

$$D = \left( \frac{n_f}{n_t} \right) \cdot 100, \quad (2)$$

где  $n_f$  – количество фертильных самок-сеголеток; штук,  $n_t$  – общее количество самок-сеголеток, штук;

3) средняя величина выводка:

$$\bar{P} = \frac{\sum_{i=1}^N E_i}{n_r}, \quad (3)$$

где  $E_i$  – количество эмбрионов и/или послеплодных пятен у  $i$ -й самки, штук;  $n_r$  – количество обследованных самок, штук.

*Климатические данные* получены из базы данных по работе метеостанций в период проведения учетов красной полевки в исследуемом регионе по температуре и осадкам [20]. В качестве климатических признаков для моделей использовали следующие показатели, которые могут оказывать влияние на выживаемость животных:

- 1) среднемесячная температура окружающего воздуха в мае, °С;
- 2) среднемесячная температура окружающего воздуха в сентябре (заморозки и возвращение тепла приводят к намоканию животных и последующему охлаждению, а также формированию ледяной корки, препятствующей добыванию корма), °С;
- 3) среднемесячная величина осадков в мае, мм;
- 4) среднемесячная величина осадков в сентябре (дожди и мокрый снег при низкой температуре воздуха потенциально губительны и приводят к переохлаждению животных), мм;
- 5) средняя температура под снегом в январе-апреле текущего года и в октябре-декабре предыдущего года непосредственно в местах обитания животных в холодный период, °С. Показатель рассчитали исходя из данных температуры окружающего воздуха и температуры под снегом, полученных с помощью датчиков-логгеров, которые собирали данные непрерывно с октября по апрель в пункте «Буюнда» в течение 9 лет [21]. Использовали температуру воздуха и уровень снега как предикторы для расчета температуры под снегом и создали экземпляр модели линейной регрессии:

$$T_{ss} = \omega_0 + \omega_1 \cdot T_a + \omega_2 \cdot S, \quad (4)$$

где  $T_{ss}$  – расчетная температура под снегом, °С;  $\omega_0$  – свободный член (интерсепт),  $\omega_1$  – коэффициент для температуры воздуха;  $\omega_2$  – коэффициент для уровня снежного покрова;  $T_a$  – температура окружающего воздуха, °С;  $S$  – уровень снежного покрова, см. Подставив рассчитанные коэффициенты, уравнение принимает вид:  $T_{ss} = - 0.39 + 0.53 T_a - 0.001 S$ , что позволяет предсказать температуру под снегом на основе известных значений температуры воздуха и уровня снежного покрова.

*Подготовка датасета.* Для подготовки наборов данных, их обработки и машинного обучения использовали язык программирования Python (библиотеки Pandas, Numpy, Scikit-learn [22]). Вычислительные эксперименты проводили в рабочей среде Google Colaboratory (Google LLC).

В целях улучшения обобщающей способности модели, устойчивости к вариациям в данных, проведена аугментация за счет создания новых примеров на основе существующих, путем добавление небольшого случайного шума (стандартное отклонение шума 10 % от стандартного отклонения значений исходных данных). Размер данных в каждом наборе увеличен в пять раз.

Так исходный набор данных  $X = \{x_1, x_2, \dots, x_n\}$ , содержащий  $n$  элементов,  $x_i$  – отдельный элемент исходного набора данных, где индекс  $i$  меняется от 1 до  $n$ . Дополнительные варианты элемента  $x_i$  созданы по формуле:

$$x_i^{(l)} = x_i + N(0, \sigma_N), \quad l = 1, 2, 3, 4, \quad i = 1, 2, \dots, n, \quad (5)$$

где  $x_i^{(l)}$  – дополнительный вариант исходного элемента  $x_i$  с добавлением шума;  $N(0, \sigma_N)$  – случайный шум, генерируемый из нормального распределения с нулевым средним и стандартным отклонением  $\sigma_N$ . Индекс  $l$ , проходящий значения от 1 до 4, создает четыре модификации для каждого исходного элемента. Полный набор данных после аугментации имеет следующую структуру  $X' = \{x_1, x_2, \dots, x_n, x_1^{(1)}, x_1^{(2)}, x_1^{(3)}, x_1^{(4)}, \dots, x_n^{(1)}, x_n^{(2)}, x_n^{(3)}, x_n^{(4)}\}$ .

Так как характеристики входного набора данных измерялись в разных единицах, поэтому данные стандартизировали:

$$z = \frac{x - \mu}{\sigma}, \quad (6)$$

где  $z$  – стандартизированное значение признака;  $x$  – исходное значение признака,  $\mu$  – среднее значение признака в исходном наборе данных,  $\sigma$  – стандартное отклонение признака в исходном наборе данных.

Анализовали три набора аугментированных и стандартизированных данных: в целом по региону (далее – Регион) – все пункты на трансекте «Магадан»-«Анадырь» ( $n = 180$ ) и отдельно по пунктам с наиболее протяженными рядами исследований – пункт «Буюнда» ( $n = 45$ ) (далее – Колыма) и пункт «Челомджа» ( $n = 50$ ) (далее – Приохотье). Представленные наборы данных позволили провести сравнительный анализ производительности двух моделей машинного обучения: случайного леса и многослойного персептрона на основе демографических (состояние популяции красной полевки) и климатических параметров (состояние среды обитания) из двух географических пунктов и в целом по Региону.

**Таблица 2.** Перечень признаков и их описание для моделей прогнозирования численности популяций красной полевки

Переменная	Наименование признака	Сокращение/ транслитерация
$x_1$	относительная численность животных в августе года	abundance_fact
$x_2$	доля размножающихся самок-сеголеток	puberty
$x_3$	величина выводка	litter_size
$x_4$	средняя температура окружающего воздуха в мае	may_temp
$x_5$	средняя температура окружающего воздуха в сентябре	sep_temp
$x_6$	средний уровень жидких осадков в мае	may_precip
$x_7$	средний уровень жидких осадков в сентябре	sep_precip
$x_8$	средняя температура под снегом в январе-апреле текущего года и в октябре-декабре предыдущего года	subsnow_temp
$y$	относительная численность животных в августе текущего года	abundance_pred

*Отбор признаков для моделей.* Выбор информативных и наиболее важных признаков (факторов) является определяющим шагом для разработки эффективных алгоритмов машинного обучения. В работе каждый набор данных за один год исследований, содержащий наблюдения за состоянием популяций красной полевки и средой их обитания, представлен как вектор признаков, но без привязки к временной последовательности. Если традиционный временной ряд прогнозирует будущие значения на основе последовательности прошлых значений, то предсказание по векторам рассматривает каждое наблюдение как отдельную точку в пространстве признаков, формируя векторные представления данных. Вектор признаков для каждого года включал климатические и демографические данные

текущего года (предикторы) и целевой признак (таргет) – относительную численность популяции красной полевки на год ( , ) (табл. 2).

Анализ на мультиколлинеарность проведен на основе матрицы коэффициентов ранговой корреляции Спирмена:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \quad (7)$$

где  $d_i$  – разница между ранжировками соответствующих пар объектов по двум признакам (ранги пары  $x_i$  и  $y_i$ ),  $n$  – количество пар объектов.

Дополнительно, для обнаружения мультиколлинеарности признаков рассчитывали фактор инфляции дисперсии (Variance Inflation Factor, VIF), который используется для количественного измерения степени мультиколлинеарности среди предикторов (признаков) в регрессионной модели. Он рассчитывается отдельно для каждого признака и отражает, насколько увеличилась дисперсия оценочного коэффициента. Фактор инфляции дисперсии для  $i$ -го признака вычисляли по следующей формуле:

$$VIF_i = \frac{1}{1 - R_i^2}, \quad (8)$$

где  $R_i^2$  – коэффициент детерминации. Учитывали, что если  $VIF_i > 5$ , значит высока вероятность мультиколлинеарности, если  $VIF_i > 10$ , то такие признаки нельзя использовать в модели.

Вычисление важности признаков проводили с помощью метода перестановки – пермутации, который эффективен для нелинейных («древесных») или непрозрачных (нейросетевых) моделей и включает в себя случайное перемешивание значений одного признака и наблюдение за результирующим ухудшением оценки модели. Проводилась оценка и сравнение нормализованной пермутационной важности:

$$ImportanceNorm(i) = \frac{I_i}{\sigma(I_i)}, \quad (9)$$

где  $I_i$  – абсолютная важность  $i$ -го признака, вычисленная как среднее снижение качества модели после перемешивания,  $\sigma(I_i)$  – стандартное отклонение значений важности  $i$ -го признака, получаемых при многократных перемешиваниях. Пусть  $Q_{orig}$  – исходное качество модели (до перемешивания),  $Q_{perm}^{(k)}$  – качество модели после  $K$ -го перемешивания  $i$ -го признака;  $K$  – количество повторений (в нашем случае  $K = 30$ ). Тогда абсолютная важность  $i$ -го признака равна среднему изменению качества модели:

$$I_i = \frac{1}{K} \sum_{k=1}^K (Q_{orig} - Q_{perm}^{(k)}). \quad (10)$$

Таким образом, стандартное отклонение важности признака:

$$\sigma(I_i) = \sqrt{\frac{1}{K-1} \sum_{k=1}^K [Q_{\text{orig}} - Q_{\text{perm}}^{(k)} - I_i]^2}. \quad (11)$$

*Модели прогнозирования.* В категории обучения с учителем решали задачу регрессии для прогнозирования относительной численности популяций красной полевки на один год вперед используя модели случайного леса и многослойного персептрона (предсказание значения  $y(t_n + 1)$  при заданной последовательности  $y(t_1), y(t_2), \dots, y(t_n)$ ). В наборах данных выделяли обучающую (для непосредственного обучения модели), валидационную (для контроля качества модели в процессе обучения и предотвращения переобучения) и тестовую (оценка качества готовой модели) выборки. Разделение производили в следующих пропорциях. Первоначально набор данных делили на две части: обучающая выборка (80 %), тестовая выборка (20 %). Затем обучающая выборка дополнительно была разделена ещё на две части: тренировочную выборку (75 % от первоначальной обучающей выборки) и валидационную выборку (25 % от первоначальной обучающей выборки). Валидационная выборка также использовалась для настройки гиперпараметров.

Проведены вычислительные эксперименты на двух моделях (случайный лес и многослойный персептрон) на трёх наборах данных (Регион, Колыма, Приохотье).

Гиперпараметры настраивали с помощью методов оптимизации автоматически через поиск по сетке (Grid Search) на валидационной выборке.

Пусть дана модель  $M$  с множеством гиперпараметров  $\Theta = (\theta_1, \theta_2, \dots, \theta_K)$ . Пространство возможных комбинаций гиперпараметров задали решёткой значений:

$$\Theta = \{\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(N)}\}, \quad (12)$$

где  $N$  – общее количество уникальных комбинаций гиперпараметров.

Исходный набор данных разделяли на пять частей (фолдов), затем для каждой комбинации гиперпараметров выполняли кросс-валидацию. Оставляли один фолд как тестовый, а остальные четыре фолда объединяют в тренировочную выборку. Далее обучали модель на тренировочной выборке с гиперпараметрами  $\theta$  и затем оценивали на тестовом фолде. Итоговая оценка для каждой комбинации гиперпараметров – это среднее значение метрики качества по всем пяти фолдам.

Лучший набор гиперпараметров  $\hat{\theta}$  выбирали, когда он показывал высшее качество на кросс-валидации:

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \frac{1}{K} \sum_{k=1}^K J(M, \theta, k), \quad (13)$$

где  $J(M, \theta, k)$  – оценка качества модели  $M$  с гиперпараметрами  $\theta$  на  $k$ -м фолде.

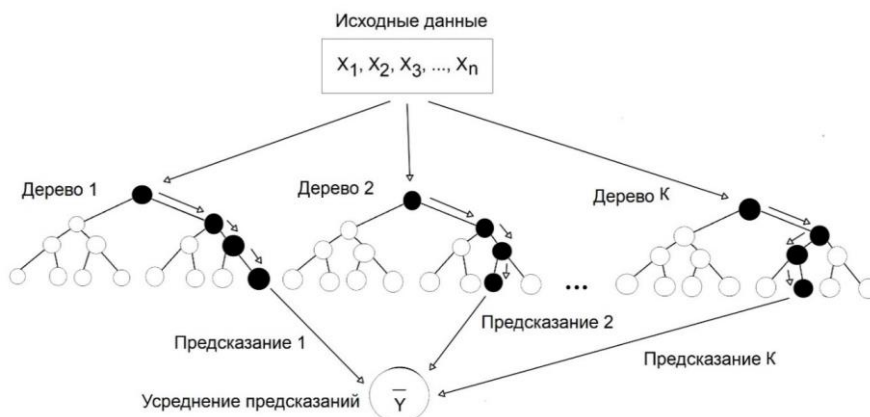


Рис. 1. Схема регрессионной модели случайного леса

Случайный лес – модель на основе ансамблевого подхода, которая объединяет множество деревьев решений для улучшения точности и устойчивости модели (рис. 1). Каждое дерево в лесу делает предсказание, а итоговое предсказание случайного леса формируется на основе усреднения предсказаний всех деревьев. Так имеется пространство признаков  $X$ , целевая переменная  $Y$ , и ансамбль из  $B$  деревьев решений  $\{T_b(x)\}_{b=1}^B$ , тогда итоговое предсказание случайного леса строится как среднее предсказаний всех деревьев:

$$F(x) = \frac{1}{B} \sum_{b=1}^B T_b(x), \quad (14)$$

где  $T_b(x)$  – предсказание  $b$ -го дерева для объекта  $x$ .

Полная формула случайного леса может быть представлена как

$$F(x|B, d, m) = \frac{1}{B} \sum_{b=1}^B T_b(x| \theta_b), \quad (15)$$

где  $T_b(x| \theta_b)$  – отдельное дерево решений, построенное с гиперпараметрами  $\theta_b = (B, d, m)$ ;  $B$  – количество деревьев;  $d$  – максимальная глубина дерева;  $m$  – минимальное количество образцов для разделения узла.

Конфигурации гиперпараметров для настройки случайного леса:  $B \in \{50, 100, 200\}$ ;  $d \in \{\text{"none"}, 10, 20, 30\}$  («none» соответствует неограниченной глубине);  $m \in \{2, 5, 10\}$ .

Многослойный персептрон – это нейронная сеть, которая используется для обучения сложных нелинейных зависимостей в данных. Она состоит из входного слоя, одного или нескольких скрытых слоев и выходного слоя (в модели регрессии – один нейрон) (рис. 2). Каждый нейрон в слое соединён со всеми нейронами следующего слоя через веса, которые определяют силу влияния одного нейрона на другой. Передача сигнала между слоями происходит следующим образом: входной слой принимает данные, далее сигнал проходит через скрытые слои, где каждая активация вычисляется как композиция операции умножения матриц и функции активации, а выходной слой выдаёт итоговое предсказание.

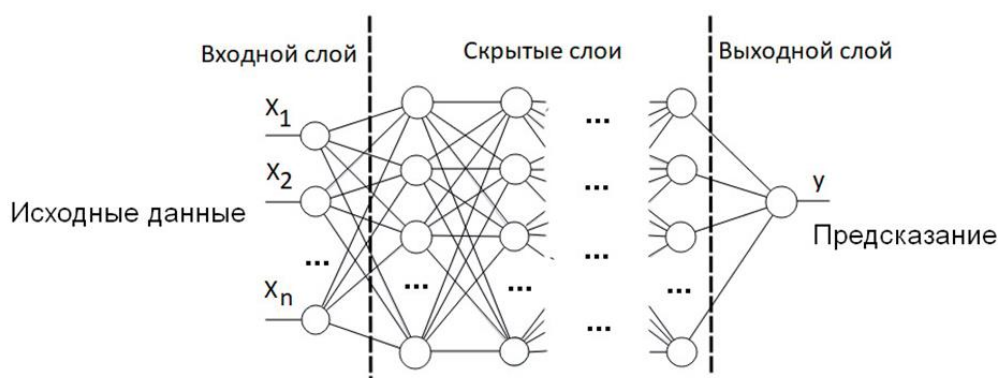


Рис. 2. Схема регрессионной модели многослойного персептрона

Передача сигнала между слоями через композицию операций матриц:

$$z^{(l)} = \omega^{(l)} a^{(l-1)} + b^{(l)}, \quad (16)$$

где  $z^{(l)}$  – активация текущего слоя  $l$ ;  $\omega^{(l)}$  – веса связей текущего слоя;  $a^{(l-1)}$  – вектор активации предыдущего слоя;  $b^{(l)}$  – смещающие члены.

Скрытые слои используют активационные функции выпрямленный линейный блок (Rectified Linear Unit, ReLU):

$$\sigma(z) = \max(0, z), \quad (17)$$

где  $\sigma(z)$  — выходное значение функции активации;  $z$  — входное значение нейрона.

Оптимизацию весов проводили методом стохастического градиентного спуска с функцией потерь, которая минимизирует разницу между предсказанными и действительными значениями. Шаг обновления весов  $\omega$ :

$$\omega^{t+1} = \omega^t - \eta \nabla_{\omega} L(\omega), \quad (18)$$

где  $\omega^t$  — текущие веса на итерации  $t$ ,  $\nabla_{\omega} L(\omega)$  — градиент функции потерь по весам,  $\eta$  — скорость обучения.

Веса — это параметры модели, которые определяют, насколько сильно каждый входной сигнал влияет на выходной сигнал нейрона. Так в многослойном персептроне с 8 входными нейронами, 5 нейронами в первом скрытом слое и 3 нейронами во втором скрытом слое веса между слоями будут выглядеть следующим образом:

– веса между входным слоем и первым скрытым слоем: матрица размером  $8 \times 58 \times 5$ , где каждый элемент  $\omega_{ij}$  представляет вес связи между  $i$ -м входным нейроном и  $j$ -м нейроном первого скрытого слоя;

– веса между первым и вторым скрытым слоями: матрица размером  $5 \times 35 \times 3$ , где каждый элемент  $\omega_{ij}$  представляет вес связи между  $i$ -м нейроном первого скрытого слоя и  $j$ -м нейроном второго скрытого слоя;

– веса между вторым скрытым слоем и выходным слоем: матрица размером  $3 \times 13 \times 1$ , где каждый элемент  $\omega_{ij}$  представляет вес связи между  $i$ -м нейроном второго скрытого слоя и выходным нейроном.

Регуляризация Тихонова (также известная как регуляризация L2) используется в машинном обучении для предотвращения переобучения модели:

$$L_2 = \sum (y - \hat{y})^2 + \lambda \sum_{i=0}^d \omega_i^2, \quad (19)$$

где  $\hat{y} = \sum_{i=0}^d \omega_i x^i$ ,  $y$  — истинное значение целевой переменной,  $\hat{y}$  — предсказанное значение целевой переменной,  $\lambda$  — коэффициент регуляризации, который контролирует силу штрафа,  $\omega_i$  — значения весов для  $i$ -го признака.

Регуляризация добавляет штраф за большие значения весов модели, что помогает уменьшить сложность модели и улучшить её способность к обобщению на новые данные.

Гиперпараметры, такие как коэффициент регуляризации Тихонова L2 с целевой функцией и количество эпох обучения, подбирали экспериментально, чтобы достичь наилучших результатов.

Применяли следующие значения гиперпараметров для настройки нейросети: коэффициент регуляризации  $L_2$ :  $\lambda \in \{0.001, 0.01, 0.1\}$ , максимальное количество эпох обучения ( $T$ ):  $T \in \{1000, 2000, 3000\}$ .

*Метрики оценки моделей.* Коэффициент детерминации ( $R^2$ ), который показывает долю вариации зависимой переменной объясняемой моделью, вычисляли по формуле:

$$R^2 = 1 - \frac{SSE}{SST}, \quad (20)$$

SSE — сумма квадратов остатков (ошибок):

$$SSE = \sum_{i=1}^N (y_i - \hat{y}_i)^2, \quad (21)$$

SST — общая сумма квадратов:

$$SST = \sum_{i=1}^N (y_i - \bar{y}_i)^2, \quad (22)$$

где  $y_i$  — фактическое значение целевой переменной,  $\hat{y}_i$  — предсказанное значение целевой переменной,  $\bar{y}_i$  — среднее значение целевой переменной,  $N$  — количество наблюдений.

Среднеквадратичная ошибка (Root Mean Square Error, RMSE) показывает среднюю ошибку модели в тех же единицах, что и целевая переменная:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}. \quad (23)$$

Точность предсказательных моделей оценивали путем сравнения коэффициента детерминации ( $R^2$ ) и среднеквадратичной ошибки (RMSE), полученных на тестовых и валидационных данных. Эти метрики дают представление о том, насколько хорошо прогнозы модели соответствуют фактическим значениям. Значения RMSE удобны для сравнения, так они выражаются в тех же единицах, что и целевой признак.

## РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Прогнозирование численности мелких грызунов является сложной задачей, требующей комплексного подхода и учета множества факторов [23]. Экспертный подбор основных признаков, влияющих на численность популяции красной полевки, включал абиотические и биотические факторы. Воздействие на выживаемость красных полевок может оказывать температура под снегом в местах обитания (снежный покров оказывает теплоизоляционный эффект, и его величина компенсирует прямое воздействие отрицательной температуры воздуха), а также количество жидких осадков ранней осенью и поздней весной, когда положительная температура воздуха в течение дня сменяется отрицательной (намокание шерстного покрова и переохлаждение животных). Внутрипопуляционные факторы предотвращают губительное перенаселение при чрезмерной эксплуатации ресурсов путем снижения потенциала размножения популяции: сокращение периода размножения при высокой численности зверьков и количества выводков за сезон, удлинение периода полового созревания молодняка и отсрочка их включения в размножение, снижение средней величины выводка.

Мультиколлинеарность между переменными возникает, когда используемые признаки сильно коррелируют друг с другом, что может негативно повлиять на интерпретацию модели и её стабильность. Анализ мультиколлинеарности выполнен на корреляционной матрице, полученной на данных до проведения аугментации. Статистически значимая связь отмечена для ряда признаков, но в большинстве случаев она слабая ( $r_s < 0.5$ ) (рис. 3). Более высокие значения  $r_s$  отмечены для климатических показателей: между признаком `subsnow_temp` и признаками `sep_temp`, `may_precip` и `sep_precip` (~ 0.53–0.58) и могут быть охарактеризованы как связь умеренной силы. В целом, сильных связей ( $r_s > 0.7$ ) между признаками не отмечено.

Для более точной диагностики дополнительно провели VIF-анализ. Судя по полученным коэффициентам, модель испытывает умеренный риск мультиколлинеарности (табл. 3). Коэффициенты фактора инфляции дисперсии ни для одного из признаков не превысили принятые пороговые значения (5 или 10).

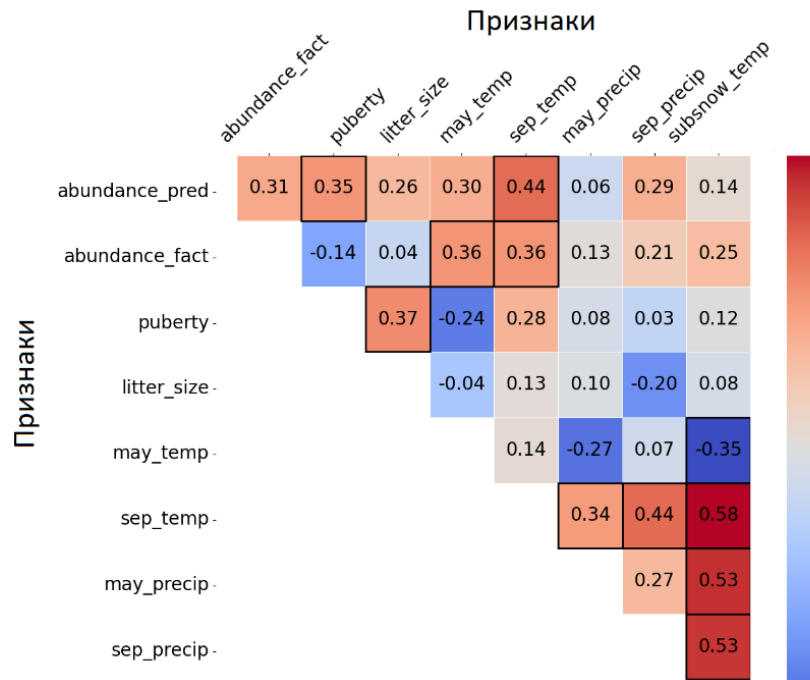


Рис. 3. «Тепловая карта» корреляций целевого признака и предикторов (внутри ячеек значение  $r_s$ , обведенные жирным ячейки – при  $p < 0.05$ )

Таблица 3. Коэффициенты фактора инфляции дисперсии для признаков

Признак	VIF
sep_temp	2.51
subsnow_temp	2.46
may_temp	1.92
sep_precip	1.65
puberty	1.44
may_precip	1.37
abundance_fact	1.36
litter_size	1.32

Случайный лес. Оптимальные гиперпараметры модели, построенной на разных наборах данных приведены в таблице 4.

Для набора данных по Региону отмечается рост значения RMSE от обучающей выборки к тестовой – модель демонстрирует признаки переобучения. В то же время, достаточно высокие значения  $R^2$  свидетельствуют, что модель хорошо объясняет вариацию данных, хотя и с некоторым снижением на валидационной и тестовой выборках.

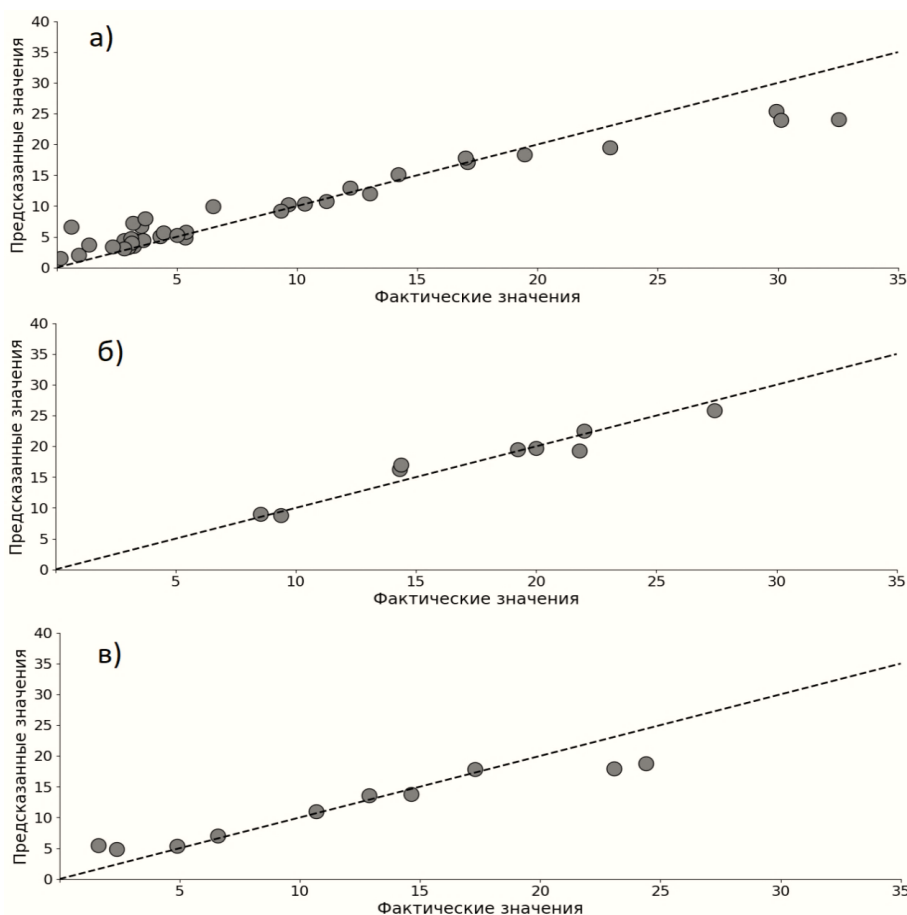
Для набора данных по пункту Колыма модель показывает хорошую обобщающую способность, так как значение  $R^2$  на валидации и тестировании близки к обучающей выборке. Этот набор данных показывает наилучшие результаты RMSE: отклонение прогноза численности по этой метрике на разных выборках минимально.

Модель на наборе данных Приохотье демонстрирует значительное снижение значений  $R^2$  на тестовой выборке, что может указывать на проблемы с обобщением. Рост RMSE на валидационной и, особенно, на тестовой выборках, указывает на переобучение модели.

**Таблица 4.** Гиперпараметры и метрики модели случайного леса на различных наборах данных

Наборы данных	Оптимальные гиперпараметры	Метрики		
		Случайная выборка	RMSE	R <sup>2</sup>
Регион	200 деревьев для формирования ансамбля; деревья в модели случайного леса могут расти до максимальной глубины; разделение узла происходит, если он содержит минимум 2 образца.	Обучение	1.56	0.97
		Валидация	2.63	0.92
		Тест	2.68	0.90
Колыма	100 деревьев для формирования ансамбля; деревья в модели случайного леса могут расти до максимальной глубины; разделение узла происходит, если он содержит минимум 2 образца.	Обучение	1.12	0.99
		Валидация	1.08	0.98
		Тест	1.52	0.93
Приохотье	100 деревьев для формирования ансамбля; деревья в модели случайного леса могут расти до максимальной глубины; разделение узла происходит, если он содержит минимум 2 образца.	Обучение	1.18	0.95
		Валидация	1.81	0.88
		Тест	2.86	0.86

Сравнение фактических и предсказанных значений численности популяций красной полевки на основе модели случайного леса представлено на рисунке 4. Отклонения наблюдаются преимущественно при высоких значениях численности. Прогноз отклоняется от фактических при показателях численности примерно на значениях от 20 экземпляров /100 ловушко-суток. Особенно это проявляется на наборе данных по Приохотью, где отмечена значительная разница между анализируемыми метриками случайных выборок.



**Рис. 4.** Фактические и предсказанные значения численности популяций красной полевки в модели случайного леса (здесь и далее – а, б, в, соответственно, Регион, Колыма и Приохотье)

Проведённый анализ важности признаков показал, что наиболее значимыми для модели случайного леса на наборе данных Регион являются такие демографические параметры как доля размножающихся самок-сеголеток и величина выводка, а для Приохотья – численность животных в предыдущем году и размер выводка. Из абиотических факторов важными оказались температура окружающего воздуха в мае и сентябре (рис. 5).

Стоит отметить, что для Региона и Колымы первостепенное значение имеет доля размножающихся самок-сеголеток, в то время как на наборе данных по Приохотью этот признак, напротив, не столь важен для прогноза. Всё это указывает на то, что значимость разных признаков (факторов), как абиотической, так и биотической природы – может в значительной мере отличаться на наборах данных, полученных из различных точек ареала вида и в разные периоды. Можно заключить, что значимость демографических параметров высока, потому как в целом по Региону (данные по 6 пунктам) скорость полового созревания молодняка и величина выводка имеют первостепенно значение для модели .

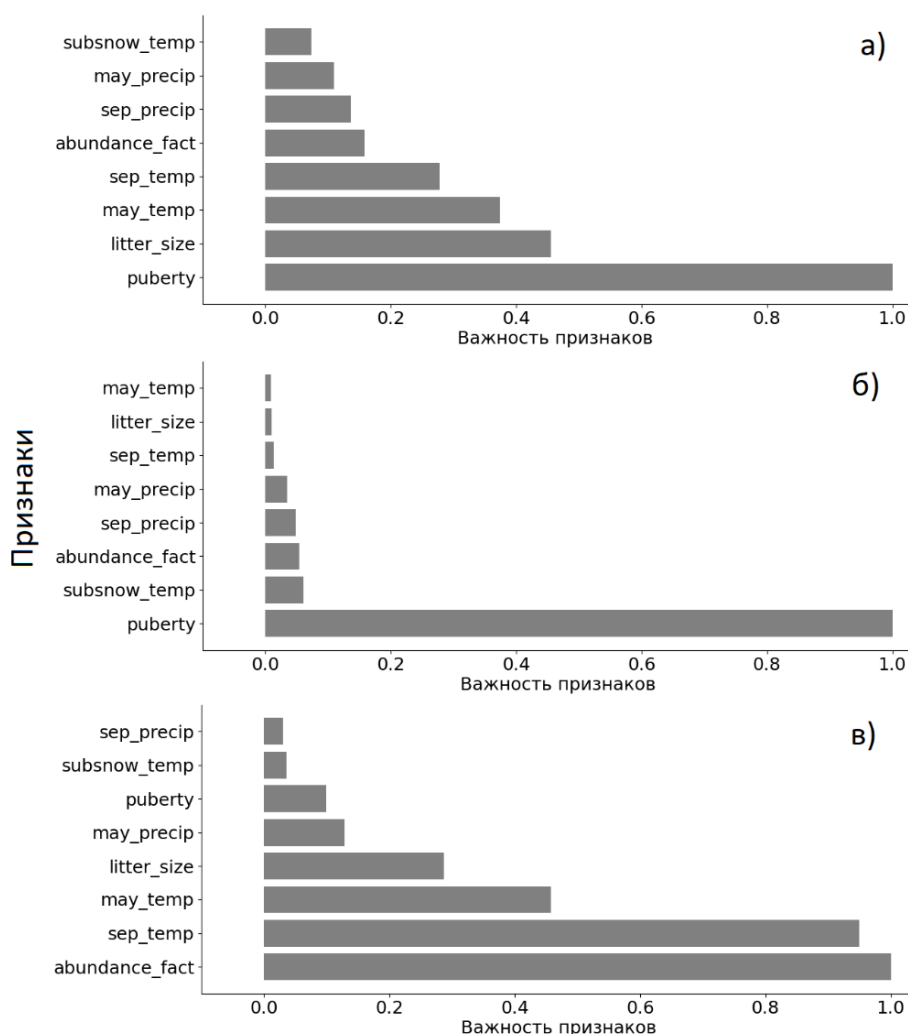


Рис. 5. Важность признаков для прогноза в модели случайного леса, установленная с помощью пермутационного анализа

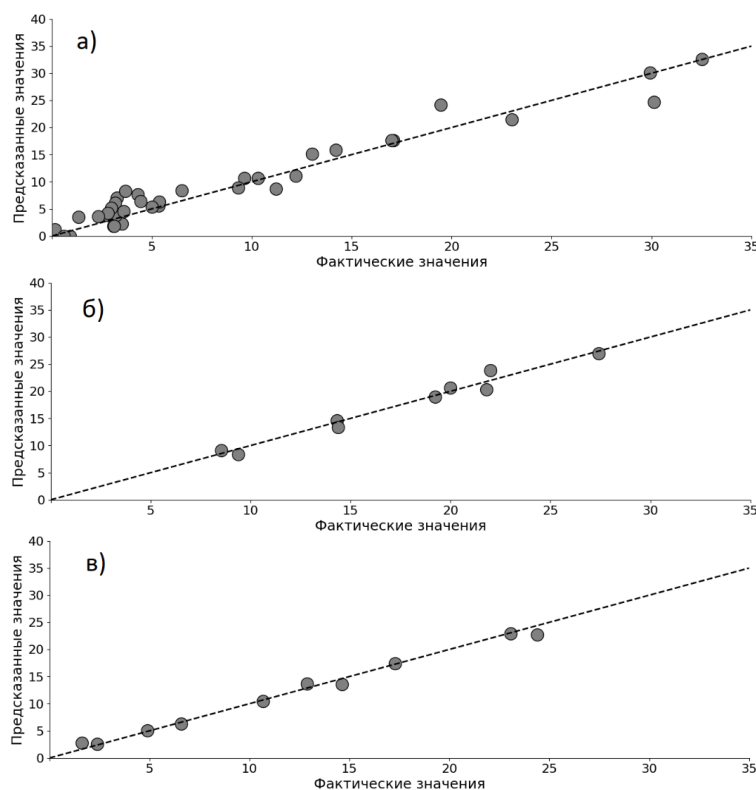
*Многослойный перцептрон.* Оптимальные гиперпараметры модели, построенной на разных наборах данных приведены в таблице 5. Для набора Регион в связи с ростом значений RMSE от обучения к тесту наблюдается значительное снижение производительности на валидационной и тестовой выборках, что указывает на переобучение модели. В то же время, высокие значения  $R^2$  показывают, что даже при

заметном снижении на валидационной и тестовой выборках модель всё же хорошо объясняет вариацию данных.

**Таблица 5.** Метрики модели многослойного персептрона на различных наборах данных

Наборы данных	Оптимальные гиперпараметры	Метрики		
		Случайная выборка	RMSE	R <sup>2</sup>
Регион	$\lambda = 0.001$ указывает на то, что модель не сильно штрафует за сложность; hidden_layer_sizes: (100, 100) – модель имеет два скрытых слоя, каждый из которых содержит 100 нейронов; max_iter: 2000 – максимальное количество итераций для обучения.	Обучение	0.53	0.99
		Валидация	2.04	0.95
		Тест	2.40	0.92
Колыма	$\lambda = 0.1$ указывает на то, что модель не слишком сильно штрафует за сложность; hidden_layer_sizes: (100, 100) – модель имеет два скрытых слоя, каждый из которых содержит 100 нейронов; max_iter: 1000 – максимальное количество итераций для обучения.	Обучение	0.68	0.99
		Валидация	1.73	0.95
		Тест	0.99	0.97
Приохотье	$\lambda = 0.1$ указывает на то, что модель не слишком сильно штрафует за сложность; hidden_layer_sizes: (150, 75) – модель имеет два скрытых слоя, каждый из которых содержит 100 нейронов; max_iter: 2000 – максимальное количество итераций для обучения.	Обучение	0.44	0.99
		Валидация	0.86	0.97
		Тест	0.77	0.99

На наборе данных Колыма по значениям RMSE на обучающей, валидационной и тестовой выборках модель показывает стабильные результаты по сравнению с Регионом (на тестовых данных RMSE < 1). По значениям R<sup>2</sup> модель также демонстрирует хорошие результаты на всех выборках, особенно на тестовой.



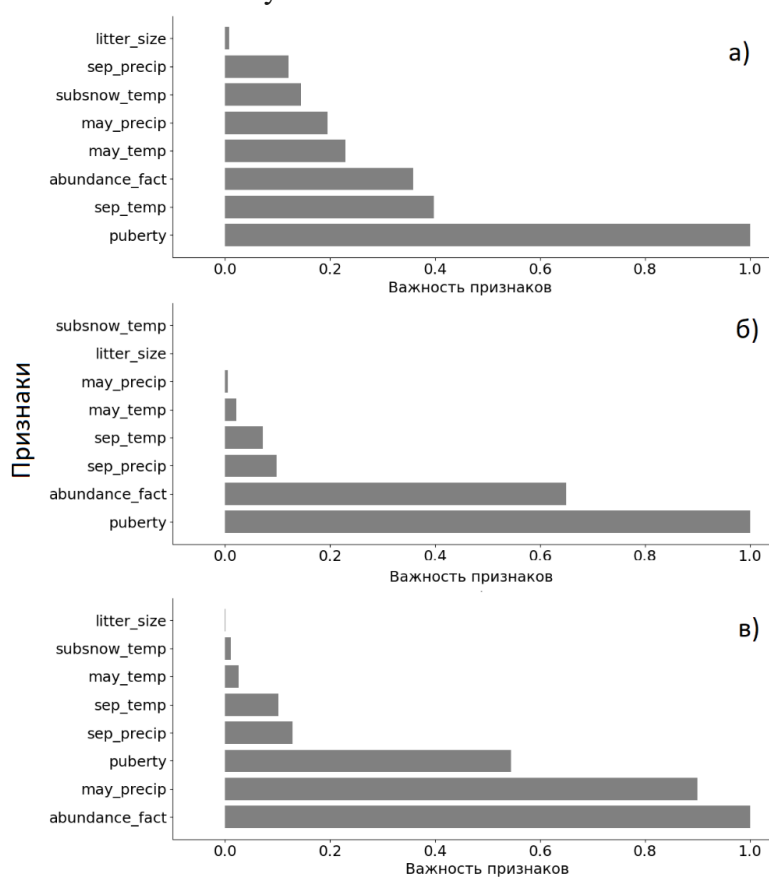
**Рис. 6.** Фактические и предсказанные значения численности популяций красной полевки в модели многослойного персептрона

По RMSE лучший результат на тестовой выборке, среди всех наборов данных, получен для Приохотья. Метрика  $R^2$  модели показывает высокую объясняющую способность (значения на обучающей и тестовой выборках равны).

Сравнение фактических и предсказанных значений на разных наборах данных представлено на рисунке 6. Прогноз достаточно точно отражает фактическую ситуацию. Полученный результат лучше, чем при использовании модели случайного леса: заметных отклонений предсказуемых и фактических значений численности популяций красной полевки не выявлено.

Анализ важности признаков показал (рис. 7), что наиболее значимыми для обсуждаемой модели по Региону являются демографические параметры (доля размножающихся самок-сеголеток и фактическая численность животных). Большое значение имеет температура осенью и весной. Осадки в мае и сентябре, а также подснежная температура оказывают меньшее влияние на прогноз.

На наборе данных по Колыме выделяется влияние демографических параметров (доля размножающихся самок-сеголеток и фактическая численность). Осадки и температуры в сентябре и мае заметно менее важны. Подснежная температура климатической зимы и величина выводка вообще не оказывают влияния на предсказание численности в этом пункте.



**Рис. 7.** Важность признаков для прогноза в модели многослойного персептрона, установленная с помощью пермутационного анализа

В пункте Приохотье также важны демографические параметры – фактическая численность животных и доля размножающихся самок-сеголеток, а из абиотических факторов заметное влияние на прогноз оказывают осадки в мае и сентябре, гораздо меньше – температура сентября и мая, и совсем слабое влияние – подснежная температура. Стоит отметить, что величина выводка не имела значения для прогноза ни на одном наборе данных.

В целом, многослойный перцептрон на тестовых выборках показывает лучшие результаты по RMSE на всех наборах данных, чем случайный лес, который демонстрирует более низкую производительность модели в предсказании. Многослойный перцептрон показывает более высокие значения  $R^2$ , что говорит о том, что эта модель лучше объясняет вариацию целевой переменной, выделяет существенные закономерности в данных, уменьшая влияние случайных факторов, и подходит для прогнозирования, так как она лучше экстраполирует имеющиеся зависимости на новые данные.

Многослойный перцептрон демонстрирует стабильные результаты по всем метрикам, что делает его более надежным выбором для всех наборов данных. Случайный лес, хотя и показывает заслуживающие внимание результаты на некоторых наборах данных (например, Колыма), в целом уступает многослойному перцептрон по значениям метрик.

Анализ важности признаков показал, что демографические параметры (доля размножающихся самок-сеголеток и численность популяции в предыдущем году) являются доминирующими факторами для прогнозирования численности.

## ЗАКЛЮЧЕНИЕ

Проведен сравнительный анализ моделей прогнозирования численности популяций красной полевки на один год вперед, построенных на основе случайного леса и многослойного перцептрона с использованием демографических и климатических данных для восточного сектора Субарктики в целом (Регион), а также для отдельных пунктов (Колыма и Приохотье).

Подбор признаков для построения моделей прогнозирования произведен оптимально: все использованные признаки – демографические и климатические параметры – несильно коррелируют между собой и имеют невысокие показатели коэффициентов VIF, что указывает на отсутствие мультиколлинеарности.

На основе анализа метрик RMSE и  $R^2$ , можно сделать вывод, что модель многослойного перцептрона превосходит модель случайного леса по эффективности на всех трех наборах данных численности популяций красной полевки. Многослойный перцептрон демонстрирует меньшую ошибку предсказания, более высокие значения коэффициента детерминации, а соответственно лучшую обобщающую способность.

Случайный лес – показывает приемлемые результаты, особенно по  $R^2$ , но менее точные по RMSE. Таким образом, для прогнозирования предпочтительнее использовать нейросетевые методы, например, многослойный перцептрон. В случае необходимости выполнения быстрого моделирования или интерпретируемости можно использовать модель случайного леса.

Анализ важности признаков показал, что в качестве предикторов для моделей прогнозирования, помимо климатических факторов, целесообразно использовать демографические показатели популяций животных.

Применение методов машинного обучения для прогнозирования численности на один год вперед показали свою эффективность, несмотря на ограниченный набор данных о состоянии популяций и условиях их обитания. Модели случайного леса и многослойного перцептрона решают задачу регрессии и предсказывают целевую переменную с высокой точностью (значения RMSE невелики по сравнению с размахом значений целевой переменной), что может применяться на практике для прогнозирования численности красной полевки и других массовых видов мелких грызунов, в соответствии с предложенным дизайном работы, для решения задач эпидемиологии и защиты растений.

## СПИСОК ЛИТЕРАТУРЫ

1. Лидикер В. Популяционная регуляция млекопитающих: эволюция взгляда. *Сибирский экологический журнал*. 1999. № 1. С. 5–13.
2. Пантелеев, П.А. Динамика численности грызунов: возможные пути к решению проблемы. *Сибирский экологический журнал*. 2008. № 1. С. 195–203.
3. Sheftel B.I. Role of different mechanisms in type determination of population dynamics for small mammals from boreal forestry zone. In: *Biological Diversity and Nature Conservation: theory and practice for teaching*. М.: KMK Scientific Press, 2010. P. 130–143.
4. Роговин К.А., Мошкин М.П. Авторегуляция численности в популяциях млекопитающих и стресс. Штрихи к давно написанной картине. *Журнал общей биологии* 2007. Т. 68. № 4. С. 244–267.
5. Фрисман Е.Я., Неверова Г.П., Кулаков М.П., Жигальский О.А. Явление мультирежимности в популяционной динамике животных с коротким жизненным циклом. *Доклады академии наук*. 2015. Т. 460. № 4. С. 488–493. doi: [10.7868/S0869565215040258](https://doi.org/10.7868/S0869565215040258)
6. Жигальский О.А. Анализ популяционной динамики мелких млекопитающих *Зоологический журнал*. 2002. Т. 81. № 9. С. 1078–1106.
7. Чернявский Ф.Б., Лазуткин А.Н. *Циклы леммингов и полевок на Севере*. Магадан: ИБПС ДВО РАН, 2004. 150 с.
8. Прислегина Д.А., Дубянский В.М., Платонов А.Е., Малецкая О.В. Влияние природно-климатических факторов на эпидемиологическую ситуацию по природно-очаговым инфекциям. *Инфекция и иммунитет*. 2021. Т. 11. № 5. С. 820–836. doi: [10.15789/2220-7619-EOT-1631](https://doi.org/10.15789/2220-7619-EOT-1631)
9. Lhoumeau S., Pinelo J., Borges P.A.V. Artificial Intelligence for Biodiversity: Exploring the Potential of Recurrent Neural Networks in Forecasting Arthropod Dynamics Based on Time Series. *Ecological Indicators*. 2025. V. 171. P. 113–119. doi: [10.1016/j.ecolind.2025.113119](https://doi.org/10.1016/j.ecolind.2025.113119)
10. Rammer W., Seidl R. Harnessing Deep Learning in Ecology: An Example Predicting Bark Beetle Outbreaks. *Front. Plant Sci*. 2019. V. 10. No. 1327. P. 1–9. doi: [10.3389/fpls.2019.01327](https://doi.org/10.3389/fpls.2019.01327)
11. Ceia-Hasse A., Sousa C.A., Gouveia B.R., Capinha C. Forecasting the abundance of disease vectors with deep learning. *Ecological Informatics*. 2023. V. 78. P. 1–10. doi: [10.1016/j.ecoinf.2023.102272](https://doi.org/10.1016/j.ecoinf.2023.102272)
12. Şahinarslan F.V., Tekin A.T., Çebi F. Application of machine learning algorithms for population forecasting. *Int. J. Data Science*. 2021. V. 6. No. 4. P. 257–270. doi: [10.1504/IJDS.2021.122770](https://doi.org/10.1504/IJDS.2021.122770)
13. Tanmoy F.M., Hossain Z., Tasfia O., Abrar Hamim M., Sadekur Rahman M., Tarek Habib M. Machine Learning Modeling for Population Forecasting. In: *Innovations in Electrical and Electronics Engineering. ICEEE 2024: Lecture Notes in Electrical Engineering*. Eds. Kalam A., Mekhilef S., Williamson S.S. Vol. 1295. Springer, Singapore, 2025. P. 213–228. doi: [10.1007/978-981-97-9112-5\\_13](https://doi.org/10.1007/978-981-97-9112-5_13)
14. Mutaz M., En-Bing L. Advancing Population Dynamics Analysis: Leveraging AI-Enhanced Mathematical Techniques. In: *ACMLC '24: Proceedings of the 2024 6th Asia Conference on Machine Learning and Computing*. 2025 P. 146–150. doi: [10.1145/3690771.3690797](https://doi.org/10.1145/3690771.3690797)
15. Steindorf V., Hamna Mariyam K.B., Stollenwerk N., Cevidanes A., Barandika J.F., Vazquez P., García-Pérez A.L., Aguiar M. Forecasting invasive mosquito abundance in the Basque Country, Spain using machine learning techniques. *Parasit and Vectors*. 2025. V. 18. № 109. doi: [10.1186/s13071-025-06733-y](https://doi.org/10.1186/s13071-025-06733-y)

16. Евстегнеева В.А., Честнова Т.В., Смольянинова О.Л. О нейросетевом моделировании и прогнозировании эпизоотий туляремии на территории Тульской области. *Вестник новых медицинских технологий. Электронное издание*. 2014. Т. 8. № 1. doi: [10.12737/7240](https://doi.org/10.12737/7240)
17. Евстегнеева В.А. К вопросу о математических методах прогнозирования заболеваемости природно-очаговыми инфекциями. *Вестник новых медицинских технологий. Электронное издание*. 2014. Т. 8. № 1. doi: [10.12737/7241](https://doi.org/10.12737/7241)
18. Евстегнеева В.А., Смольянинова О.Л., Логвинов С.И. Сравнительный анализ математических методов в прогнозировании заболеваемости лептоспирозом. *Успехи современной науки*. 2016. Т. 1. № 8. С. 173–179.
19. Ямборко А.В. *Популяционная экология лесных полевков (род Clethrionomys) Северо-Восточной Азии*: диссертация ... кандидата биологических наук: 03.02.08. Магадан, 2015. 202 с.
20. Веселов В.М., Прибыльская И.Р., Мирзеабасов О.А. *Специализированные массивы для климатических исследований*. URL: <http://aisori-m.meteo.ru/waisori/select.xhtml> (дата обращения: 10.07.2025).
21. Лазуткин А.Н., Ямборко А.В., Киселев С.В. Популяционная динамика лесных полевков (р. Clethrionomys) верховьев Колымы (р. Буюнда). *Вестник СВНЦ ДВО РАН*. 2012. № 4. С. 66–74.
22. *Scikit-learn – Машинное обучение в Python*. URL: <https://scikit-learn.ru/stable> (дата обращения: 09.02.2026).
23. Жигальский О.А. Анализ методов прогнозирования заболеваемости зоонозными инфекциями. *Эпидемиология и вакцинопрофилактика*. 2012. № 3 (64). С. 27–31.

Рукопись поступила в редакцию 24.10.2025, переработанный вариант поступил 30.12.2025.  
Дата опубликования 04.03.2026.

===== DATA MINING =====

## **Application of Machine Learning Models to Predict Small Rodent Populations (using *Myodes rutilus* as an example)**

**Yamborko A.V.<sup>1,2,3</sup>, Timoshilov V.I.<sup>3,4</sup>**

<sup>1</sup>*Federal Center for Analysis and Assessment of Technogenic Impact, Research Center for Rare and Endangered Species of Animals and Plants (branch), Moscow, Russia*

<sup>2</sup>*Institute of Biological Problems of the North, Far Eastern Branch of the Russian Academy of Sciences, Magadan, Russia*

<sup>3</sup>*Moscow Institute of Physics and Technology, Moscow, Russia*

<sup>4</sup>*Kursk State Medical University, Kursk, Russia*

**Abstract.** A comparative analysis of machine learning regression models was conducted to predict changes in the relative abundance of natural northern red vole populations one year in advance. Demographic and climate data for the eastern

subarctic as a whole, as well as for two locations within the region, were used to train, validate, and test random forest and multilayer perceptron models. Each dataset, containing annual observations of various indicators, is represented in the model as a feature vector (the target feature is relative abundance one year in advance; the predictors are population and climate data), without reference to a temporal structure. Traditional time series forecasting derives future values from a sequence of past values, while vector forecasting treats each observation as a separate point in feature space, forming vector representations of the data. A multilayer perceptron has been shown to yield better results and forecast accuracy across all samples. A random forest is characterized by lower robustness and accuracy. Neural network techniques, like a multilayer perceptron, function better for both the entire area and specific locations. A random forest model can be used if fast modeling or interpretability is required. Using red vole populations as an example it was shown that an accurate forecast of the relative abundance of small rodents one year in advance could be obtained using a time-limited dataset on the status of populations and their habitats. Machine learning models can be applied to solving problems in epidemiology and plant protection.

**Key words:** *machine learning, random forest, multilayer perceptron, population forecasting, *Myodes rutilus**