УДК 639.2.053.8:597.552.511(265.51) DOI: 10.15853/2072-8212.2020.59.76-96

ИСПОЛЬЗОВАНИЕ МЕТОДА RANDOM FOREST В ЦЕЛЯХ ПРОГНОЗИРОВАНИЯ ПОДХОДОВ ГОРБУШИ СЕВЕРО-ВОСТОКА КАМЧАТКИ

М.Г. Фельдман

Вед. н. с., Камчатский филиал Всероссийского научно-исследовательского института рыбного хозяйства и океанографии («КамчатНИРО»)
683000 Петропавловск-Камчатский, Набережная, 18
Тел.: 8 (4152) 41-27-01. E-mail: feldman.m.g@kamniro.ru

ПРОГНОЗИРОВАНИЕ, ГОРБУША, КЛАССИФИКАЦИЯ, МОДЕЛИРОВАНИЕ, СЛУЧАЙНЫЙ ЛЕС, ДЕРЕВО РЕШЕНИЙ, ЯЗЫК R, АНАЛИЗ ДАННЫХ

Для прогнозирования подходов камчатской горбуши используется мощный современный метод машинного обучения Random Forest — случайный лес деревьев решений. В качестве предикторов используются помесячные данные климатических индексов. В работе применяется итеративный способ отбора наиболее важных факторов. Выбор лучшей модели осуществлен по наименьшей ошибке на тестовых данных. Алгоритм применяемого метода оформлен на языке R.

USING THE DECISIONS OF THE RANDOM FOREST ALGORITHM FOR THE PURPOSES OF FORECASTING PINK SALMON RUNS ON NORTH-EASTERN KAMCHATKA

Mark H. Feldman

Leading Scientist, Kamchatka Branch of Russian Research Institute of Fisheries and Oceanography (“KamchatNIRO”)
683000 Petropavlovsk-Kamchatsky, Naberezhnaya Str., 18
Tel.: +7 (4152) 41-27-01. E-mail: feldman.m.g@kamniro.ru

FORCASTING, PINK SALMON, CLASSIFICATION, MODELING, RANDOM FOREST, CODE R, DATA MINING

Forecasting of pink salmon runs in Kamchatka uses modern powerful method of machine learning Random Forest (random forest of decision trees). Monthly data of climate indices are used as predictors. Forecasting applies iterative way of selection of the most important factors. Decision about the best model is based on the least error on test data. The algorithm of the method is written in R language.

Горбуша составляет основу добычи лососей Камчатки, поэтому важность прогнозирования ее подходов трудно переоценить. Вместе с тем задача прогнозирования именно этого вида достаточно сложна, поскольку у него нет перекрывающихся поколений (как у других тихоокеанских лососей), наличие которых заметно улучшает качество прогнозных моделей (Peterman, 1982; Haeseker et al., 2007), а численность поколений подвержена сильным флуктуациям. При этом численность каждого поколения мы рассматриваем как результат комплексного воздействия плотностной зависимости (Ricker, 1954) и абиотических факторов среды. Ранее эту задачу решали с помощью общей регрессионной модели (Фельдман, Шевляков, 2015). Однако выбор объясняющих переменных в данном случае ложился полностью на исследователя: необходимо было отобрать от трех до пяти факторов из нескольких десятков самостоятельно, с визуализацией объясняющей и зависимой переменной на диаграмме рассеивания.

Естественно, что отбор предикторов машинным способом в таком случае заметно облегчил бы анализ данных. Между тем в случае нелинейных взаимоотношений предиктора и зависимой переменной даже машинный поиск простых корреляций мало что даст. Поэтому в данной работе мы, во-первых, упростили данные по зависимой переменной (подходу горбуши), придав им три уровня численности (низкий, средний и высокий), о чем подробнее будет сказано ниже, а во-вторых, использовали ансамблевый метод анализа данных, тем более что таковые были рекомендованы Межинститутской рабочей группой по методологии оценки сырьевой базы рыболовства при ФГБНУ ВНИРО (РГМ) еще в 2017 г.

Хотя распространение ансамблевых моделей произошло лишь в последнее время, их предыстория уже достаточно давняя. Так, в 1906 г. на сельскохозяйственной ярмарке в Плимуте английский исследователь и член Лондонского королевского общества сэр Френсис Гальтон, интересовавший-

ся результатами селекции, столкнулся с проводимым конкурсом, где публике предлагалось за вознаграждение угадать вес быка после разделки (Galton, 1907). У сэра Гальтона, критически относящегося к концепции демократических выборов, возникла идея, что это идеальный эксперимент, который может показать несостоятельность коллективного решения, ведь каждый из участников был заинтересован только в своем выигрыше: «Среди конкурентов были мясники и фермеры, некоторые из которых были очень опытны в оценке веса скота; другие, вероятно, руководствовались той информацией, которую они могли бы получить и своими собственными фантазиями. Средний участник, вероятно, был также хорошо подготовлен для справедливой оценки разделанного быка, поскольку средний избиратель оценивает достоинства большинства политических вопросов, по которым он голосует». Однако, к изумлению Гальтона, когда он подсчитал среднее от всех прогнозов, оно оказалось лишь на 9 фунтов отличавшееся от реального веса: 1207 фунтов составил средний прогноз, а 1198 фунтов — реальный вес разделанного быка.

Об этом и других не менее поразительных случаях правильного коллективного решения можно прочитать в интересной книге «Мудрость толпы…» финансового аналитика изданий New Yorker и Wall Street Journal Джеймса Шуровьески (2007). Между тем, название этого явления как «мудрость толпы» (The Wisdom of Crowds), на наш взгляд, не такое удачное, как в терминологии сэра Гальтона — «глас народа» (Vox Populi). Дело в том, что, как отмечают оба автора, мнение каждого участника должно быть независимым от мнения других, что вряд ли случается именно в толпе.

Примечательно, что почти за полтора века до случая с сэром Гальтоном во Франции маркизом де Кондорсе было опубликовано эссе, в котором была сформулирована теорема о жюри присяжных (Condorcet, 1785 по Мюллер, 2007). Теорема гласит, что если каждый член жюри присяжных имеет независимое мнение и вероятность правильного решения больше 0,5, тогда вероятность правильного решения присяжных в целом возрастает, и с увеличением количества членов жюри стремится к единице. И наоборот: если же вероятность правильного решения у каждого из членов жюри меньше 0,5, то та же вероятность для всего жюри стремится к нулю с увеличением количества присяжных.

Естественно, что такие наблюдения не остались без внимания ученых. Помимо экспериментов по исследованию принятия решений с различными группами людей, с развитием вычислительной техники, начали формироваться и методы машинного обучения, имитирующие эффект vox populi, — ансамблевые методы моделирования, где в качестве элементарных единиц используются какие-либо простые модели, так называемые «слабые ученики». Оказалось, что в качестве таких базовых единиц достаточно хорошо проявляют себя деревья решений (Breiman et al., 1984; Quinlan, 1986). Дерево решений представляет собой алгоритм принятия решения, имеющий четко выраженную древовидную структуру. Структура дерева включает в себя корень (root) — признак, который лучше всего делит выборку на две части или ветви, узлы (nodes) — признаки нижеследующих порядков, которые также делят свою подвыборку на две части, и листьев (leafes) — принятых решений.

Алгоритм дерева решений является «жадным», он способен распознать 100% наблюдений в выборке, при этом, однако, переобучаясь. С этим недостатком можно бороться различными методами остановки обучения: глубиной дерева, предельным количеством наблюдений в узле и др. Отметим другие достоинства этой модели:

– простая и легкая интерпретация всего пошагового процесса принятия решения,

– работает с любыми типами переменных, как с числовыми (причем любого масштаба) величинами, так и категориальными.

Вместе с тем, помимо склонности к переобучению, есть и недостатки:

– деревья нестабильны, т. к. небольшое изменение во входных данных может привести к изменению их структуры;

– деревья решений часто относительно неточны, у них большая дисперсия ошибки прогноза.

Именно с помощью ансамблевых методов можно устранить такие недостатки базовой модели дерева решений, как переобучение и дисперсия ошибки. Становление алгоритма случайного леса тесно связано с появлением техники бутстрепа — размножением выборок (Efron, 1979). В данной технике новые выборки формируются из элементов исходной случайным образом. Первоначально бутстреп использовался для оценки различных статистических параметров. В конце прошлого

века Лео Брейман предложил использовать бутстреп-выборки с замещением (чтобы выборки были той же длины, как и исходная) для построения базовых моделей ансамбля (Breiman, 1996a). Такой метод был назван им бэггингом (от англ. **b**ootstrap **agg**regation). Характерной особенностью таких выборок является то, что около 37% наблюдений не используются в построении отдельного дерева вообще и такие наблюдения можно автоматически использовать для тестирования соответствующего дерева (Breiman, 1996b). Такую ошибку тестирования принято называть ошибкой out-of-bag (OOB error — ошибка вне выборки). Она подсчитывается по наблюдениям, не вошедшим в обучение каждого дерева, и усредняется для всего ансамбля в целом. Данная ошибка является состоятельной оценкой и эквивалентна ошибке на контрольных данных (Джеймс и др., 2016; Шитиков, Мастицкий, 2017).

Бэггинг деревьев решений показал снижение ошибки прогноза за счет снижения ее дисперсии, но между тем сами отдельные деревья такого ансамбля, обучающиеся с помощью одних и тех же предикторов, получаются высококоррелированными, а это несоблюдение условия о том, что мнение каждого участника голосования должно быть независимым от других мнений. Причиной этого является то, что признаки-доминанты используются в каждом дереве, а второстепенные по важности, но вместе с тем важные предикторы вытесняются. Устранить данный недостаток помог метод случайных подпространств, показанный в работе Тин Кам Хо (Ho, 1995), который в формулировке автора является способом реализации «стохастического дискриминационного» подхода к классификации, предложенного Е. Клейнберг (Kleinberg, 1990, 1996). Суть данного метода — создание ансамбля деревьев, выращенных на одной выборке, но при этом каждое дерево строится по собственному набору признаков, который, в свою очередь, формируется из исходного набора предикторов, т. е. используется тот же бутстреп, но для признаков и без замещения. Эмпирически установлено, что для эффективного решения задач на регрессию нужно брать подвыборку предикторов размером около трети от исходного набора признаков, а для задач на классификацию — квадратный корень из исходного количества признаков.

В дальнейшем Л. Брейман (Breiman, 2001) объединил оба метода в один: каждое дерево ансамбля строится на собственной выборке с замещением, и на каждом этапе расщепления дерева рассматривается только часть предикторов. Таким образом, второстепенные по важности признаки получают свой шанс быть использованными в построении дерева, и корреляция между деревьями снижается. Метод получил название Random Forest. Его уникальность заключается в том, что несмотря на бурное развитие методов машинного обучения в последние годы и на появление различных его модификаций, он до сих пор является одним из самых лучших. Так, в сравнении 179 классификаторов из 17 семейств (дискриминантный анализ, байесовский анализ, нейронные сети, машины опорных векторов, случайные леса и другие ансамбли, обобщенные линейные модели, ближайшие соседи, частичные регрессии наименьших квадратов и главных компонент, логистическая и полиномиальная регрессия, множественные сплайны адаптивной регрессии и другие методы) на 121 наборе данных, классический вариант случайного леса оказался лучшим, показывая 90% точности в 84% наборов данных (Delgado et al., 2014).

Основные достоинства модели случайного леса (Радченко, 2017):

– также как и дерево решений, может работать с любыми типами данных;

– с увеличением количества деревьев не переобучается;

– устойчив к мультиколлинеарности признаков, никакая часть модели случайного леса не подвергается воздействию коллинеарных переменных: даже если две переменные обеспечивают одинаковую чистоту дочернего узла, можно выбрать одну из них, не ухудшая качество результата;

– не чувствителен к выбросам в данных;

– не требует тщательной настройки параметров, хорошо работает «из коробки».

Минусы:

– в отличие от одного дерева, результаты случайного леса сложнее визуально интерпретировать;

– нет формальных выводов для оценки значимости переменных (p-values), однако случайный лес позволяет оценивать важность переменных по другим критериям (таким, как уменьшение индекса Джини или точности дерева при исключении признака, и др.);

– алгоритм работает хуже многих линейных методов, когда в выборке очень много разреженных признаков (например, тексты);

– в отличие от регрессии случайный лес не умеет экстраполировать (но это можно считать и плюсом, так как не будет экстремальных значений в случае попадания выброса);

– на зашумленных данных алгоритм переобучается;

– большой размер получающихся моделей, этот недостаток можно обойти, приведя вместо самой модели программный код, что в последнее время стало принятым в научных статьях.

В данной работе мы ставим целью апробировать на данных по горбуше Северо-Восточной Камчатки совершенно новый для прогнозирования рыбных запасов метод случайного леса решающих деревьев Random Forest, хорошо зарекомендовавший себя в решении задач как на регрессию, так и на классификацию (Breiman, 2001).

Для осуществления поставленной цели решались следующие задачи:

– выбор объясняющих переменных;

– отбор значимых признаков с помощью фильтрации: построение моделей случайного леса итеративно с исключением незначимых и малозначимых предикторов на каждой итерации;

– выбор наиболее значимых признаков;

– верификация результата с помощью сторонних методов отбора признаков в случайном лесе;

– построение моделей Random Forest;

– выработка прогноза по оптимальной модели;

– оформление программного кода на получившем широкое распространение в биологической статистике языке R для воспроизводимости результатов.

Научная новизна данной работы состоит в том, что впервые в прогнозировании рыбных запасов используется метод случайного леса Random Forest. Также предложенный итеративный способ исключения незначимых предикторов для случайного леса дает лучшую сходимость окончательной модели по сравнению с другими методами исключения.

Практическая значимость работы заключается в значительном облегчении процедуры отбора объясняющих переменных, кроме того, можно рассматривать достаточно большое количество признаков, до нескольких десятков. За счет этого и автоматического тестирования модели на наблюдениях out-of-bag заметно уменьшается время принятия решения. Также программный код может использоваться в решении подобных задач для других запасов горбуши и иных видов рыб.

МАТЕРИАЛ И МЕТОДИКА

Исходные данные. В качестве зависимой переменной использованы уровни стратифицированной модели типа запас–пополнение, примененной к данным по количеству родителей и потомков горбуши Северо-Восточной Камчатки (Фельдман и др., 2018).

Различные уровни воспроизводства горбуши показаны отдельными маркерами и цветами, при этом каждый уровень описывается своей кривой зависимости пополнения R от родительского запаса S (рис. 1). Два наблюдения высокого уровня и три наблюдения низкого (с метками года нереста) не входили в обучение соответствующих кривых, и предположительно принадлежат к экстремальным классам. Но так как таких наблюдений мало

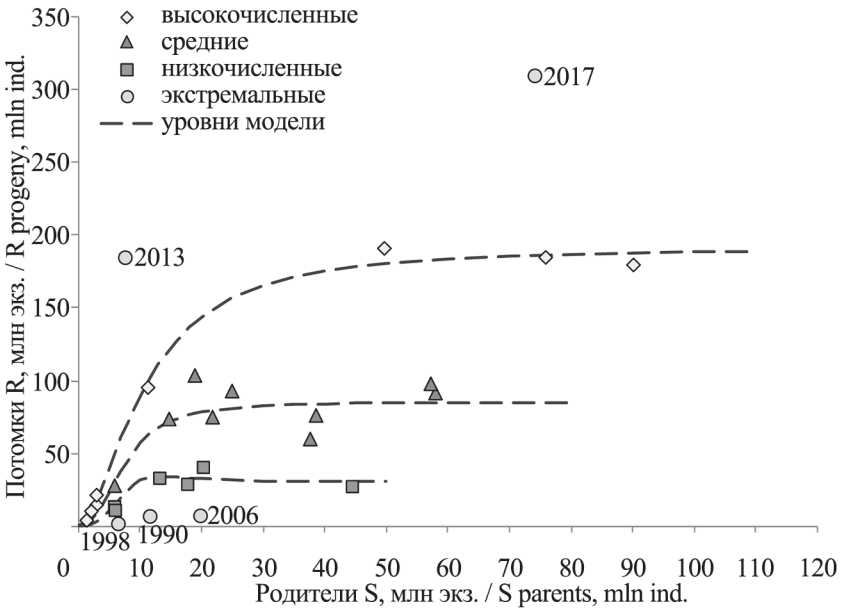

Рис. 1. Стратифицированная модель запас–пополнение для данных по горбуше северо-востока Камчатки за период 1990–2017 гг.
Fig. 1. Stratified model “parental stock – recruitment” for the data on pink salmon of North-Eastern Kamchatka for the period 1990–2017

относительно других классов, они причислены к ближайшим стратам. Таким образом, имеем три страты (класса) с почти одинаковым числом наблюдений в каждой в период с 1990 по 2017 гг. включительно: по 9 наблюдений низкого и среднего уровня и 10 наблюдений высокого уровня.

Удовлетворительно описать зависимость потомков от родителей только с помощью модели запас–пополнение не представляется возможным. Так, сравнительно одинаковые родительские запасы в различные годы формируют достаточно дифференцированное количество потомков. К примеру, уровень пропуска родителей порядка 7 млн экз. обеспечил в 2000 г. возврат 2 млн потомков, а в 2015 г. — 157 млн потомков, с разницей почти в 80 раз. Мы предполагаем, что наличие сильных флуктуаций в подходах горбуши Северо-Восточной Камчатки обусловлено преимущественно воздействием внешних факторов, о чем сообщалось ранее (Фельдман, Шевляков, 2015). Поэтому в качестве предикторов были выбраны климатические индексы, которые могут отображать климатические процессы в этой части Камчатского полуострова и к тому же имеются в свободном доступе на открытых интернет-ресурсах.

Среди таких индексов были отобраны:

– Индекс тихоокеанской декадной осцилляции (PDO), корреляции которого с изменчивостью вылова тихоокеанских лососей часто обсуждаются специалистами по рыбным запасам (Hare, Francis, 1995; Hare, 1996; Mantua et al., 1997; Mantua, Hare, 2002; Кляшторин, Лябушин, 2000; Бугаев и др., 2018). Данный индекс отображает изменения температуры поверхностных вод Северной Пацифики (севернее 20° с. ш.), которые в свою очередь могут оказывать влияние на выживаемость популяций горбуши как в течение их морского периода жизни, так и, опосредованно, во время их пресноводного периода. Надо сказать, что исходя из того, что тихоокеанские лососи имеют в течение онтогенеза критические стадии (Маркевич, Виленская, 1998; Карпенко, 1998; Шунтов, Темных, 2005; Шунтов, Темных, 2011), во время которых они могут быть более чувствительны к внешним воздействия, мы использовали помесячные данные этого и других индексов, начиная с января года нереста родительского поколения, заканчивая декабрем второго года жизни, т. е. периодом начала зимовки молоди горбуши в открытом океане. Таким образом, всего получается 24 признака за два года. Следует учесть, что индекс PDO меняется медленно, поэтому смежные помесячные данные сильно коррелируют друг с другом. Однако, как сказано выше, выбранный метод случайного леса устойчив к проблеме мультиколлинеарности. Данные индекса PDO использованы с сайта Национального управления океанических и атмосферных исследований (NOAA, США): https://www.ncdc.noaa.gov/teleconnections/pdo/.

– Индекс западно-тихоокеанского паттерна (WP), отображающий циклоническую активность в северо-западной части Тихого океана преимущественно над Камчатским полуостровом; кроме того, он тесно связан с изменчивостью краевых ледовых зон арктических морей и зимнего гидроклимата вплоть до тихоокеанского побережья Америки (Linkin, Nigam, 2008). Ранее уже были показаны значительные корреляции некоторых месячных значений WP с выживаемостью горбуши Северо-Восточной Камчатки (Фельдман, Шевляков, 2015). В этот раз с помощью случайного леса в качестве предикторов мы рассмотрим более широкий диапазон помесячных данных WP за тот же период, что и для данных по индексу PDO (источник — сайт NOAA: https://www.cpc.ncep.noaa.gov/data/teledoc/wp.shtml).

– Аналогично предыдущим использовались и помесячные данные индекса арктической осцилляции (AO) (Thompson, Wallace, 1998). Этот индекс характеризует аномалию давления в Арктике по сравнению с более южными широтами. Предположительно этот показатель может иметь значение для северо-камчатского региона, где могут иметь место прорывы холодных воздушных масс из Арктики. Данные использованы также с сайта NOAA: https://www.cpc.ncep.noaa.gov/products/precip/CWlink/daily_ao_index/ao.shtml.

Всего получилось достаточно большое количество признаков — 72. Соответственно, ввели их обозначения, где на первом месте литерное название климатического индекса, на втором числовой индекс месяца с 1 по 24. Понятно, что первые 6 месяцев (январь–июнь) для каждого индекса не совпадают по срокам жизни поколения, если считать, что нерест проходит преимущественно в июле и августе. Однако мы решили включить их, учитывая, что климатические факторы могут влиять на гидрологический режим рек весной и иметь отсроченное воздействие. Из данных признаков мы позволили себе объединить путем усреднения в один

признак данные PDO за декабрь и январь первого года жизни (обозначение PDO12_13), а также данные по индексу WP за май–сентябрь второго года жизни (обозначение WP17_21). Данные факторы уже несколько лет используются по методологии общей регрессионной модели в подготовке прогнозов подхода камчатской горбуши (Фельдман, Шевляков, 2015) и показали свою достаточно высокую значимость. Первый из них соответствует зимовальному периоду онтогенеза, второй — времени покатной миграции молоди и ее раннему морскому периоду жизни. Общая таблица предикторов и зависимой переменной имеется в репозитории https:/github.com/MarkFeld/RandomSalmon.git.

В отличие от регрессионного анализа, где корреляции можно отобрать на предварительном этапе разведки, в задаче на классификацию важны не столько корреляции, а само расположение классов в зависимости от предикторов, и отобрать предикторы на этом этапе достаточно тяжело, учитывая взаимодействия между предикторами. Хорошим для анализа с помощью случайного леса будет расположение наблюдений, если наблюдения одного класса располагаются кучно; это позволит классификатору хорошо распознать данные этих страт. Так, рассматривая на диаграмме рассеивания взаимоотношение натурального логарифма индекса выживаемости (отношения потомков к родителям) и признака PDO12-13, можно легко заметить, что наблюдения, отнесенные к низкому уровню численности (low), расположены преимущественно слева в области отрицательных значений признака, а наблюдения среднего уровня (average) — в положительной области (рис. 2А).

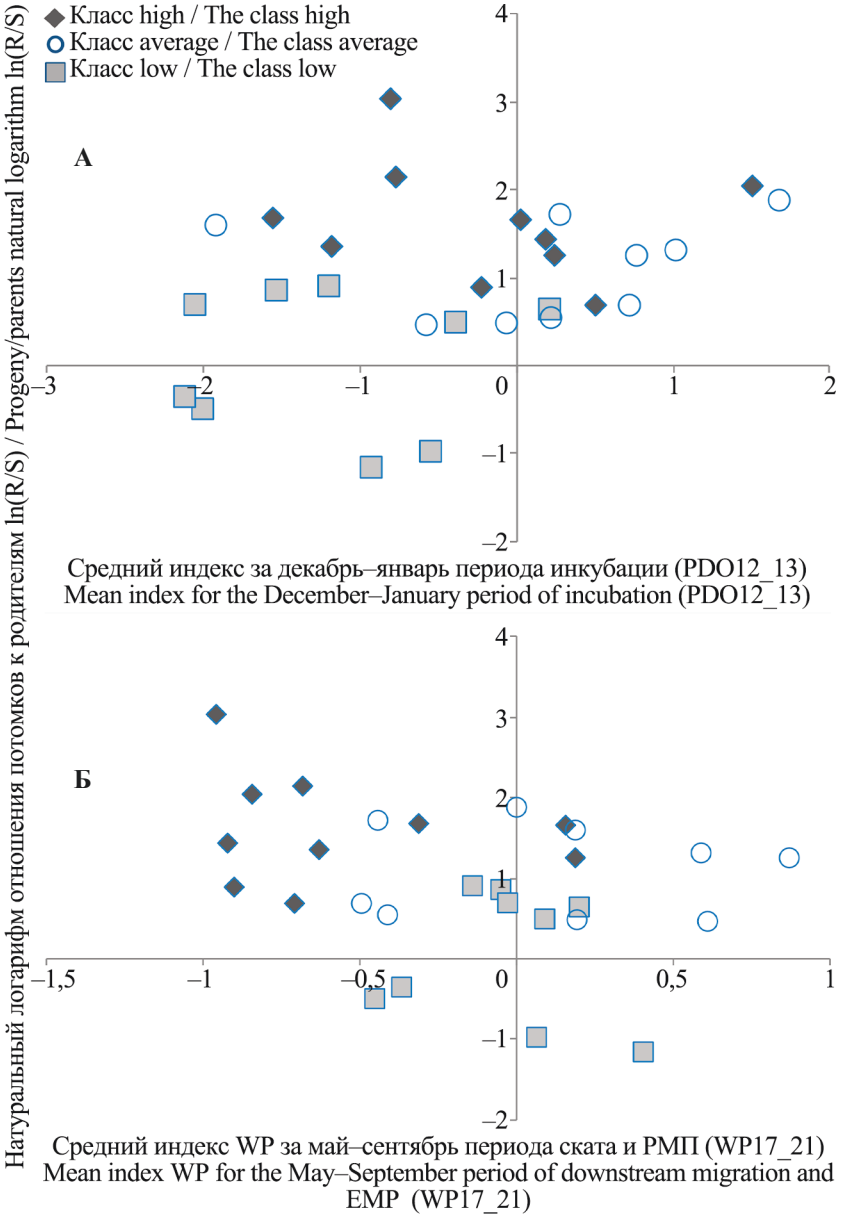

Рис. 2. Взаимодействие натурального логарифма индекса выживаемости (отношения родителей и потомков) и климатических предикторов PDO12-13 (А, период зимовки предличинок) и WP17_21 (Б, скат молоди и ранний морской период жизни). Оттенок маркеров соответствуют цветам уровней на рис. 1

Fig. 2. The correlation between the natural logarithm of the survival index (parents/progeny) and the climate predictors PDO12-13 (A, period of prolarval wintering) and WP17_21 (Б, juvenile downstream migration and early marine period of life). The colores of the markers correspond to the colors of the levels in Fig. 1

Аналогично для признака WP17_21 большая часть наблюдений высокого уровня (high) при WP17_21 < –0,5, а наблюдения класса low в средней части при –0,5 < WP17_21 < 0,5 (рис. 2Б). С помощью этих двух факторов можно обучить простое и достаточно сильное дерево решений (рис. 3).

Кроме климатических признаков, были включены еще два. Первый, натуральный логарифм численности производителей (обозначен как ln_S), согласно теории запас–пополнение (Рикер, 1954) должен оказывать непосредственное влияние на численность потомства. Согласно модельным кривым, показанным выше (рис. 1), его влияние будет значительным на численность потомства при небольших значениях числа родителей. Второй признак является коэффициентом упитанности производителей горбуши р. Хайлюли, одной из реперных рек, по которой имеются многолетние наблюдения (обозначен как k). Предположительно, данный признак будет отображать качество родительского запаса, что может оказать влияние и на потомство.

МАТЕРИАЛ И МЕТОДИКА

В итоге всего в исходном фрейме данных 70 переменных (69 предикторов и одна зависимая переменная) и 28 наблюдений с 1990 по 2017 годы. Процедуру исключения незначимых и малозначимые признаков (фильтрации) проводили итеративно. На каждой итерации строилась модель случайного леса с параметрами:

– количество признаков на каждом этапе расщепления дерева: $mtry \approx \sqrt{N}$, где N — количество предикторов;

– минимальное допустимое количество наблюдений в терминальном узле дерева (листе): $nodesize \approx \ln(N) - 1$;

– количество деревьев в лесу постоянно *ntree* = = 150;

Соответственно, на первых итерациях, при большом количестве признаков, строились «слабые» деревья небольшой глубины. После построения леса признаки сравнивались по такому параметру важности, как среднему уменьшению точности модели при их исключении — mean decrease accuracy, данный параметр вычисляется автоматически в библиотеке randomForest языка программирования R. Далее устанавливался критерий, по которому проводили исключение незначимых предикторов. В качестве такового использовали квантиль 5% — признаки с важностью менее данного критерия удалялись из фрейма данных. Затем по оставшимся признакам модель леса строилась заново и цикл повторялся. По мере убывания количества предикторов, увеличивалась разрешающая способность деревьев, обусловленная параметром nodesize. Итерации завершались, если количество признаков редуцировалось до одного предиктора (тривиальной модели). В итоге получали вектор результатов, например ошибки OOB, которые записывались на каждой итерации цикла. Выбор наилучшей модели осуществлялся по наименьшей ошибке OOB.

Первоначально, во время настройки кода, использовался только данный цикл, но так как было обнаружено, что при различных зернах генератора случайных чисел (ГСЛ), который обуславливает состав бутстреп-выборок наблюдений и необходим для воспроизводимости результатов, возникают существенные различия в моделях леса, было решено рассмотреть 1000 выборок с различными зернами ГСЛ от 1 до 1000 с помощью внешнего цикла (рис. 4). К тому же критерий, по которому удалялись малозначимые признаки, создает векторы результатов с равной длиной независимо от зерна ГСЛ. Иными словами, на соответствующих итерациях внутреннего цикла удаляется одинаковое количество признаков при любом зерне

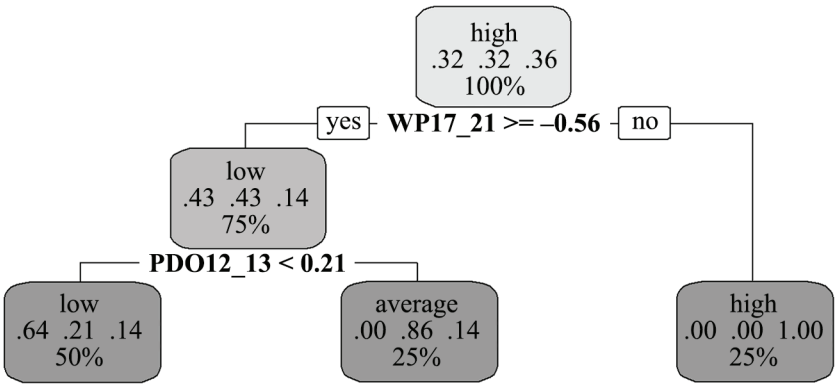

Рис. 3. Дерево решений с признаками WP17_21 и PDO12-13
Fig. 3. The tree of decisions with the traits WP17_21 and PDO12-13

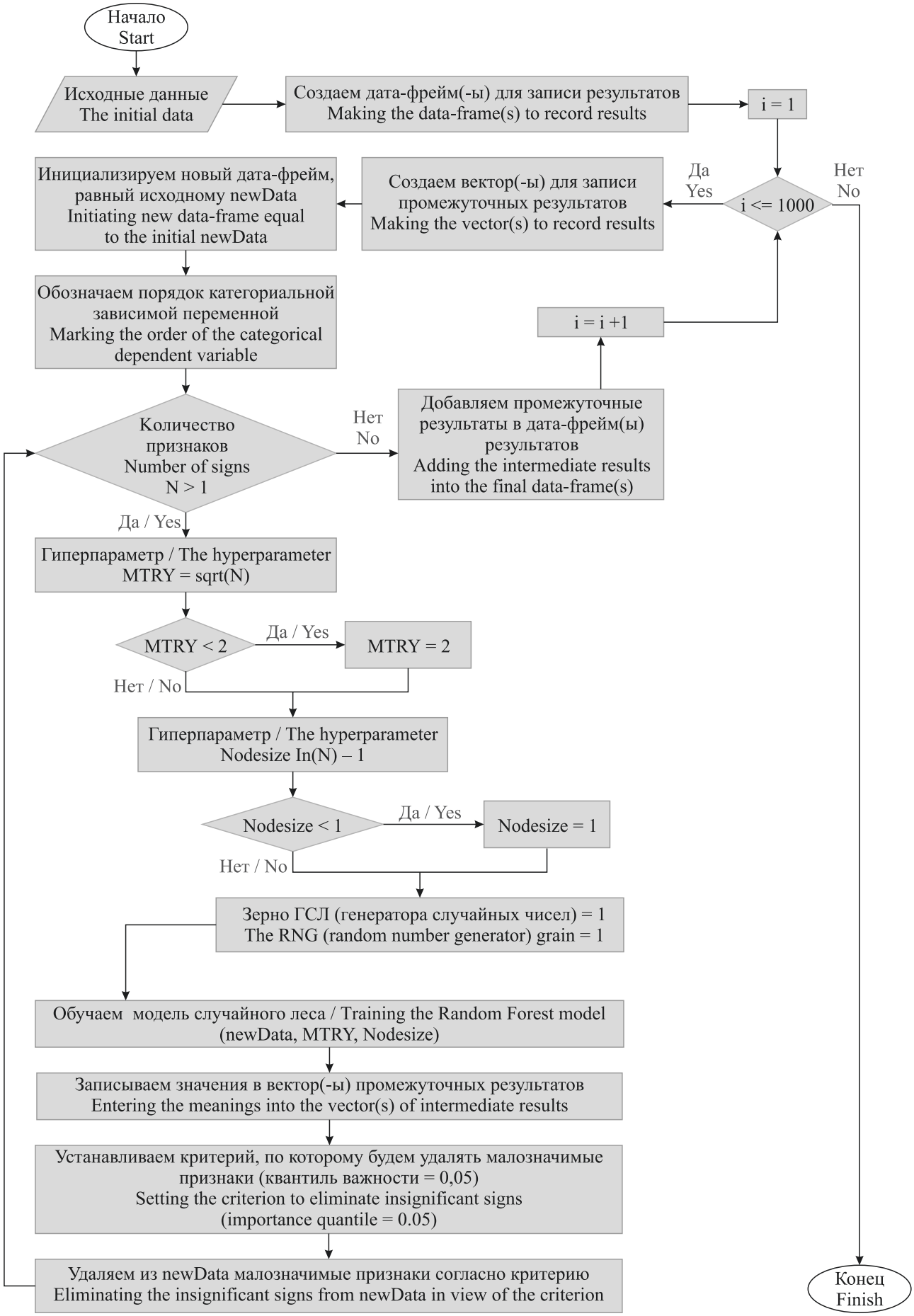

Рис. 4. Блок-схема основного алгоритма работы / Fig. 4. The block-scheem of the basis algorithm of the work

ГСЛ. Всего было получено 37 000 моделей случайного леса (1000 итераций внешнего цикла × 37 итераций вложенного цикла). Основной программный код на языке R представлен в Приложении (стр. 92) и в тексте статьи. Обработка результатов и их графическое представление проводились в среде R и MS Excel.

В среде R использовались следующие библиотеки:

– readxl — импорт данных из MS Excel в R;

– dplyr — манипуляции с данными;

– rpart.plot — графическое представление дерева решений;

– randomForest — обучение модели случайного леса;

– randomForestExplainer — анализ признаков случайного леса и взаимодействий между ними;

– Boruta — анализ и верификация признаков случайного леса;

– xlsx — экспорт данных из R в MS Excel.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Основные результаты изменения ошибки OOB для всех лесов представлены на рис. 5. Там же показаны соответствующие каждой итерации внутреннего цикла параметры MTRY и Nodesize, а также количество признаков N, участвующих в обучении. В начальных итерациях вложенного цикла количество случайных предикторов слишком велико и лес переобучается (overfitting), доля ошибок OOB в среднем около 60%, с размахом 36–82%. По мере исключения незначимых предикторов на каждой итерации, доля ошибок постепенно снижается, пока не достигает минимума (34-я итерация, 5 предикторов, доля ошибок OOB — 18%, с размахом 4–36%). Если количество признаков уменьшать дальше, то будут исключаться уже значимые признаки и модель недообучается (underfitting). Такая *U*-образная форма кривой ошибки является универсальной и справедлива практически для любых алгоритмов и методов создания предсказательных моделей (Шитиков, Мастицкий, 2017). Интересно, что в отличие от регрессионного анализа, где принято сначала брать минимум предикторов, а потом по мере необходимости их добавлять, здесь мы идем обратным путем.

Между тем, в зависимости от зерна ГСЛ (которое будет определять состав бутстреп-выборок)

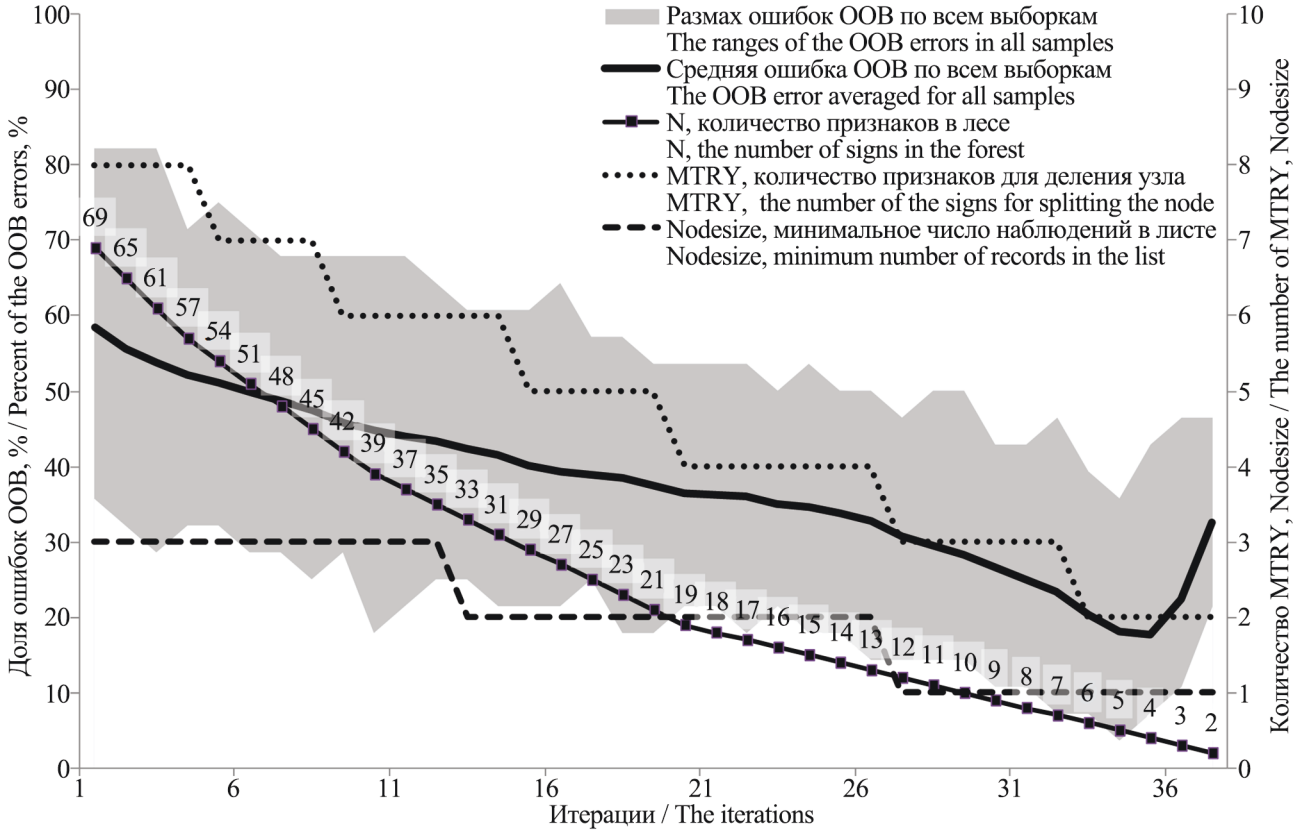

Рис. 5. Динамика изменения гиперпараметров MTRY, Nodesize и доли ошибок OOB по мере исключения наименее значимых предикторов на каждой итерации
Fig. 5. The dynamics of the changes of the hyperparameters MTRY, Nodesize and of the part of the OOB errors on eliminating the least significant predictors at each iteration

возникают различные варианты случайного леса. Мы определили размах доли ошибок OOB на каждой итерации по тысяче выборок с различными зернами ГСЛ. Минимум ошибки OOB достигнут на 34-й итерации в двух выборках (зерно ГСЛ — 853 и 968, 5 предикторов, MTRY = 2 предиктора, Nodesize = 1 наблюдение). Надо понимать, что для модели случайного леса с максимальной ошибкой на данной итерации количество предикторов то же, но их состав отличается, т. е. обучение леса идет по неверному пути, и вместо незначимого предиктора был исключен менее значимый на более ранней итерации. Как правило, для лесов с большой ошибкой (18% и выше) на 34-й итерации (условно «плохих моделей») минимумы ошибки будут встречаться либо на более ранних итерациях, когда еще не исключены значимые признаки, либо на более поздних итерациях, когда незначимые признаки уже исключены (рис. 6). Так, для «плохих» лесов на 33-й итерации ошибочно удаляется один из основных признаков, и вместо него остается какой-либо другой незначимый предиктор, который на следующей, 34-й итерации обуславливает существенный рост ошибки OOB. Далее, на 35-й итерации этот случайный предиктор удаляется, и ошибка вновь уменьшается, график изменения ошибки OOB в этой области становится W-образным. Таким образом, если исследователь будет обучать лес на неопределенном зерне ГСЛ, и он окажется «плохим», то минимальная ошибка OOB составит не более 25%, а в среднем 14%. Что можно считать достаточно хорошим результатом.

Более подробно данные результаты можно рассмотреть на гистограммах встречаемости признаков на 34-й итерации в каждой группе ошибок (рис. 7). Всего распределение составило 10 групп от минимальной ошибки OOB до максимальной (порядок на рис. 7 сверху вниз, справа налево). В первой группе, где ошибка минимальна, в обоих моделях леса участвуют 5 предикторов (WP17_21, WP7, AO5, PDO12_13 и ln_S), причем три из них (WP17_21, AO5 и ln_S) в обоих случаях находятся на первом, третьем и пятом месте по важности (в качестве меры важности использована mean deacrease accuracy). Признаки WP7 и PDO12_13 меняют свое положение: в одной модели WP7 на второй позиции, PDO12_13 на четвертой, во второй модели — наоборот. В следующей группе с ошибкой OOB в 7% на последней, пятой позиции появляется новый предиктор AO1 (35% случаев), происходит также некоторое «размытие» пяти ведущих предикторов по позициям. Далее в последующих группах появляются все новые малозначимые предикторы (обозначены оттенками серого на гистограммах), а размытие основных признаков по позициям усиливается.

В группах со средней ошибкой 32% и 35% полностью вытесняется признак ln_S. В последней группе с максимальной ошибкой OOB в 35% вытесняется также признак PDO12_13. Интересно, что ведущий признак WP17_21 удерживается на первой позиции не менее, чем в 80% случаев в любой группе. По итогам анализа ошибок OOB можно заключить, что наилучшие результаты достигаются моделью случайного леса с 5 признаками по убыванию значимости: WP17_21, WP7, AO5, PDO12_13 и ln_S. Помимо предикторов WP17_21, PDO12_13 и ln_S, которые мы априори определили как значи-

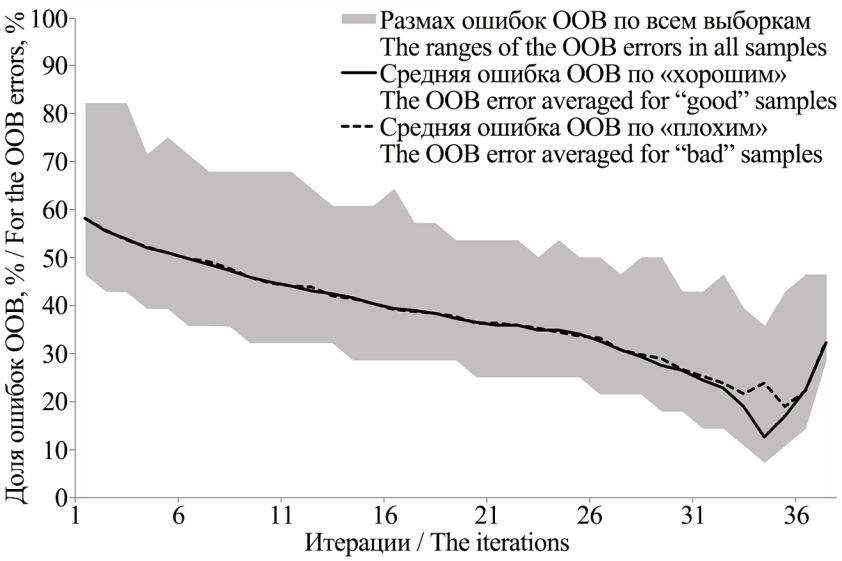

Рис. 6. Изменение средней ошибки для «хороших» выборок (ошибка OOB <18% на 34-й итерации) и «плохих» выборок (ошибка OOB на 34-й итерации ≥18%)
Fig. 6. Changing the mean error for “good” samples (the OOB error at the 34th iteration<18%) and “bad” samples (the OOB error at the 34th iteration ≥18%)

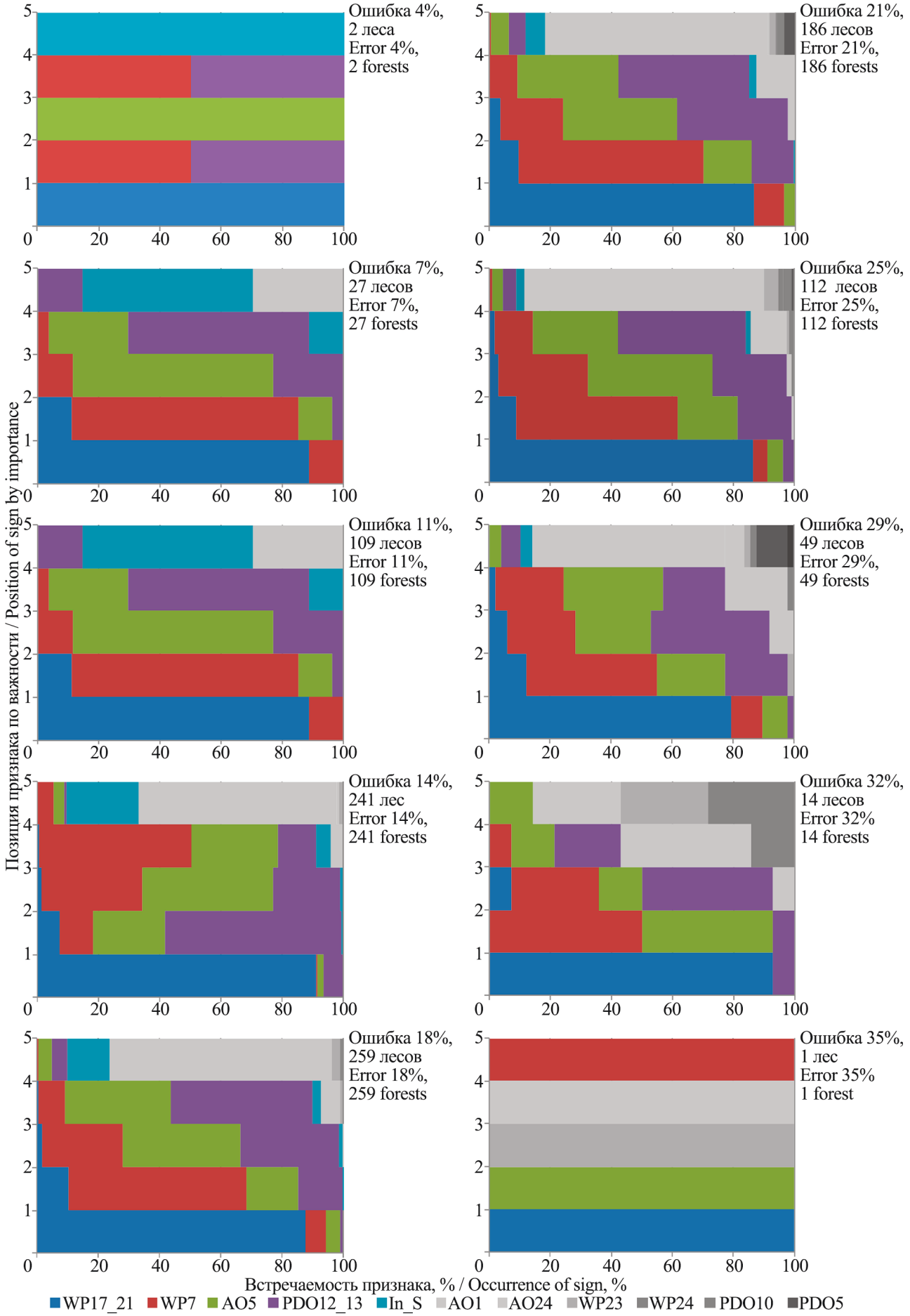

Рис. 7. Встречаемость пяти ведущих признаков из выборок с минимальной ошибкой OOB (в цвете, второстепенные признаки показаны оттенками серого) в группировках с различными уровнями данной ошибки на 34-й итерации
Fig. 7. The occurrence of five major signs in the samples with the minimal OOB error (the minor signs marked in shades of grey) in the groups with different level of the error at the 34th iteration

мые, алгоритм выявил еще два достаточно значимых фактора: WP7, AO5. Судя по числовым индексам, оба они имеют отношение к периоду нереста, причем если индекс циклонической активности за июль WP7 непосредственно совпадает с началом нерестового хода, то индекс арктической осцилляции за май (AO5) имеет скорее всего какое-то отложенное воздействие. Неожиданным явилось то, что фактор качества производителей (коэффициент упитанности k) оказался одним из худших предикторов. На рассмотренных нескольких вариантах леса данный признак исключался уже на первой итерации. Вероятно, что здесь имеет место неправильный выбор данных: коэффициент упитанности производителей для р. Хайлюля, возможно, слабо коррелирует с таковым коэффициентом для рыбы всего Северо-Востока Камчатки.

Рассмотрим теперь варианты леса с наилучшими результатами. Наименьшая ошибка OOB на 34-й итерации в 4% (в абсолютном выражении — одна ошибка из 28 наблюдений), достигнута для двух лесов с зернами ГСЛ 853 и 968 (скрипт 1).

Изменение доли ошибок OOB для каждого из классов, в зависимости от числа деревьев в случайном лесу, стабилизируется достаточно быстро — после 85 деревьев (рис. 8). Классы low и average на тестовых деревьях определены безошибочно, в классе high одна ошибка.

С помощью пакета randomForestExplainer (см. описание — Paluszyńska, электронный ресурс) в среде R можно сравнить признаки выбранного леса по различным параметрам их важности (скрипт 2).

Важность признака WP17_21 с большим отрывом больше по двум характеристикам: уменьшении индекса Джини и уменьшении точности модели. Встречаемость в узлах у признаков WP17_21, AO5, PDO12_13 и ln_S примерно на одном уровне, а признак WP7 встречается меньше всего, но в то же время по показателю уменьшения точности модели он занимает второе место после WP17_21.

Пакет также помогает визуализировать основные взаимодействия между признаками. Четыре ведущих взаимодействия показаны на рис. 9. Так, графика показывает, что взаимодействие признаков WP17_21 и PDO12_13 достаточно хорошо определяет классы low и average, а признаков WP7 и WP17_21 — классы average и high.

Проверить релевантность полученных пяти признаков можно с помощью метода Boruta из

Скрипт 1

```
  ## создаём дата-фрейм с выбранными пятью признаками и зависимой переменной
resultData <- data.frame(PDO12_13 = allData$PDO12_13,
                         WP7 = allData$WP7,
                         WP17_21 = allData$WP17_21,
                         AO5 = allData$AO5,
                         ln_S = allData$ln_S,
                         strata = allData$strata)
  ## объявляем зависимую переменную фактором
resultData$strata <- factor(resultData$strata)
  ## обозначаем порядок фактора (low < average < high)
resultData$strata <- ordered(resultData$strata,
                             levels = c("low", "average", "high"))
  ## подключаем библиотеку
library(randomForest)
  ## устанавливаем зерно ГСЛ для воспроизводимости результатов
set.seed(853)
  ## создаём объект случайного леса
resultData.rf <- randomForest(strata~ .,  xtest=NULL, ytest=NULL,
                              data=resultData, mtry=2, nodesize=1,
                              ntree=150,
                              replace=TRUE, importance=TRUE, localImp=TRUE,
                              proximity=TRUE)
  ## вывод графика ошибки OOB
plot(resultData.rf)
```

одноименного пакета (Kursa, Rudnicki, 2010; Kursa, 2020). Суть метода заключается в том, что из каждого признака создается теневой признак путем перемешивания наблюдений исходного (обозначены как Nonsense1, ...2, ...). Теневые признаки добавляются в исходную таблицу данных. Такая вдвое расширенная таблица используется для итеративного построения объектов случайного леса. Алгоритм вычисляет средний максимальный Z-балл важности для теневых признаков, обозначен в библиотеке как shadowMax (Z-балл — количество стандартных отклонений от средней важности по всем признакам, как исходных, так и теневых). На каждой итерации алгоритм сравнивает Z-баллы теневых и исходных признаков, чтобы определить, выше ли Z-балл действительного признака, чем оценка shadowMax. Если выше, то алгоритм пометит признак как подтвержденный. Такое сравнение признаков с теневыми признаками, априори незначимыми, повышает надежность оценки их важности. В результате своей работы алгоритм Boruta подтвердил важность всех пяти признаков (скрипт 3, рис. 10).

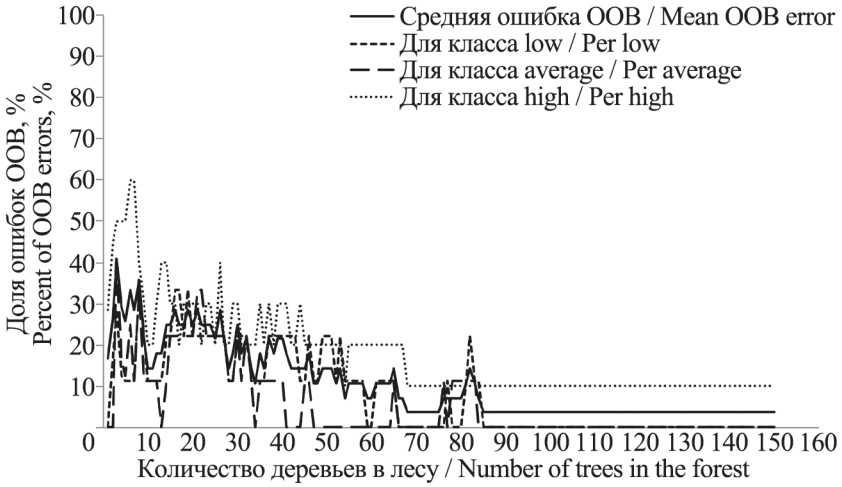

Рис. 8. Снижение доли ошибки out-of-bag в зависимости от количества деревьев в случайном лесу для каждого класса и в среднем по классам
Fig. 8. Reducing the out-of-bag error percent depending the number of trees in random forest for every class and averaged

Скрипт 2

```
 ## подключаем библиотеку
library(randomForestExplainer)
  ## проводим основную функцию пакета, результат которой выводится в файл HTML
explain_forest(resultData.rf, interactions = TRUE)
  ## код графиков основных взаимодействий
plot_predict_interaction(resultData.rf, resultData, “WP17_21”, “PDO12_13”, grid = 100,
                         main = paste0(“Взаимодействие предикторов “,
                         paste0(“WP17_21”, paste0(“ и “, “PDO12_13”))), time = NULL)
plot_predict_interaction(resultData.rf, resultData, “WP17_21”, “ln_S”, grid = 100,
                         main = paste0(“Взаимодействие предикторов “,
                         paste0(“WP17_21”, paste0(“ и “, “ln_S”))), time = NULL)
plot_predict_interaction(resultData.rf, resultData, “WP7”, “WP17_21”, grid = 100,
                         main = paste0(“Взаимодействие предикторов “,
                         paste0(“WP7”, paste0(“ и “, “WP17_21”))), time = NULL)
plot_predict_interaction(resultData.rf, resultData, “PDO12_13”, “AO5”, grid = 100,
                         main = paste0(“Взаимодействие предикторов “,
                         paste0(“PDO12_13”, paste0(“ и “, “AO5”))), time = NULL)
  ## создаем дата-фрейм характеристик важности
importance_frame <- measure_importance(resultData.rf)
  ## по выбранным характеристикам важности (с минимальными корреляциями)
  ## выводим многоплановый график по трем характеристикам важности
plot_multi_way_importance(importance_frame, x_measure = “accuracy_decrease”,
                          y_measure = “gini_decrease”,
                          size_measure = “no_of_nodes”, no_of_labels = 5)
```

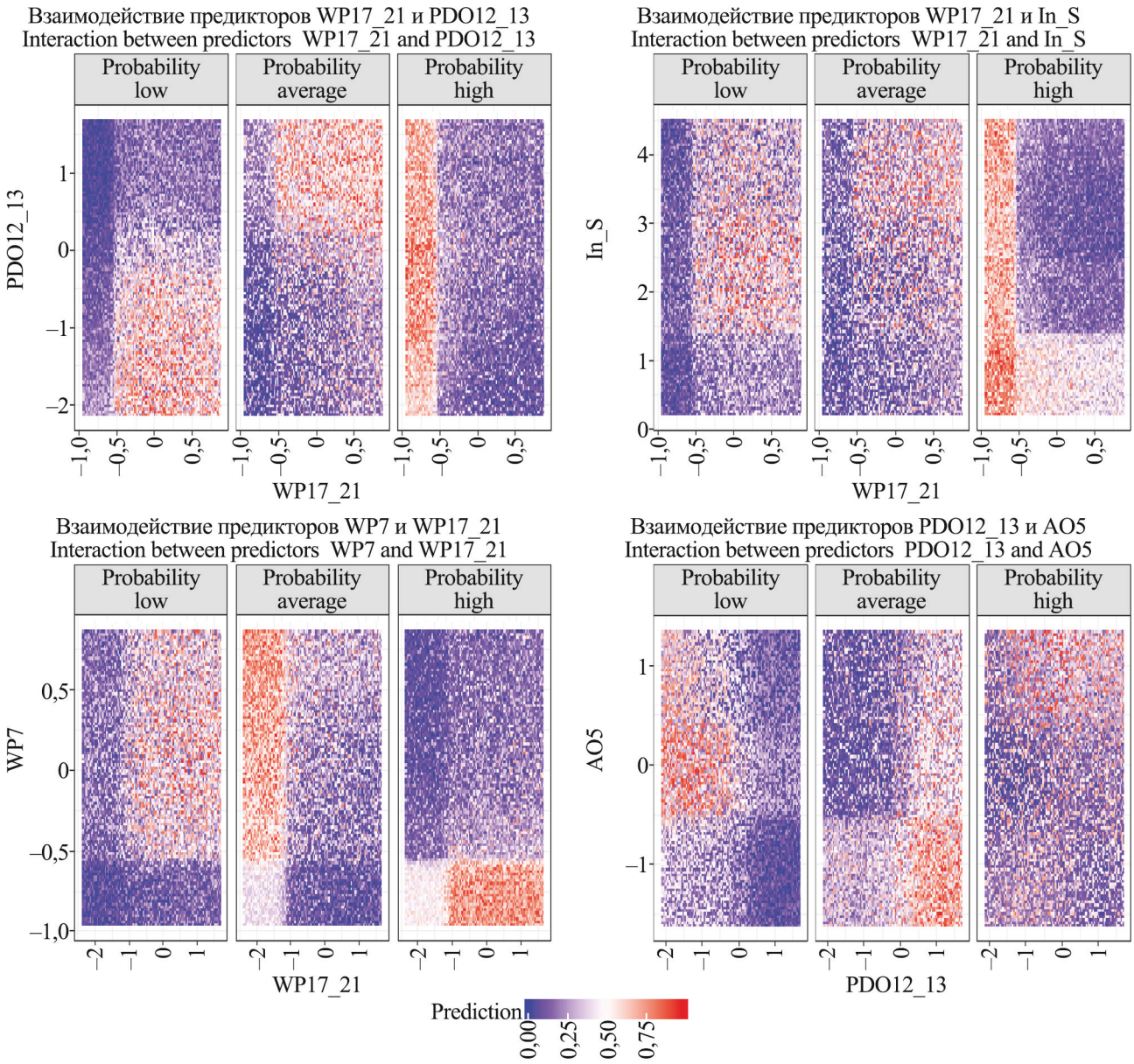

Рис. 9. Четыре основных взаимодействия между предикторами. По оси абсцисс — корневая переменная для поддерева взаимодействия
Fig. 9. Four general interactions between the predictors. The root variable for the interaction subtree is on the abscissa

Скрипт 3

```
 ## алгоритм Boruta - проверка признаков на релевантность
  ## создаём фрейм с выбранными пятью признаками и зависимой переменной
resultData <- data.frame(PDO12_13 = allData$PDO12_13,
                         WP7 = allData$WP7,
                         WP17_21 = allData$WP17_21,
                         AO5 = allData$AO5,
                         ln_S = allData$ln_S,
                         strata = allData$strata)
  ## подключаем библиотеки
library(Boruta)
library(ranger)
  ## устанавливаем зерно ГСЛ для воспроизводимости результатов
set.seed(1)
```

```
 ## создаем дата-фрейм с теневые признаки
resultData.extended <- data.frame(resultData, apply(resultData[,-6],2,sample))
 ## даем имена теневым признакам
names(resultData.extended)[7:11] <- paste("Nonsense",1:5,sep="")

 ## создаем объект Boruta
Boruta.resultData.extended <- Boruta(strata~.,data=resultData.extended,
                                     mcAdj = TRUE,
                                     doTrace=2,
                                     ntree = 150,
                                     maxRuns = 300)
 ## результат
Boruta.resultData.extended
Boruta performed 83 iterations in 14.40983 secs.
 5 attributes confirmed important: AO5, ln_S, PDO12_13, WP17_21, WP7;
 5 attributes confirmed unimportant: Nonsense1, Nonsense2, Nonsense3, Nonsense4, Nonsense5;
```

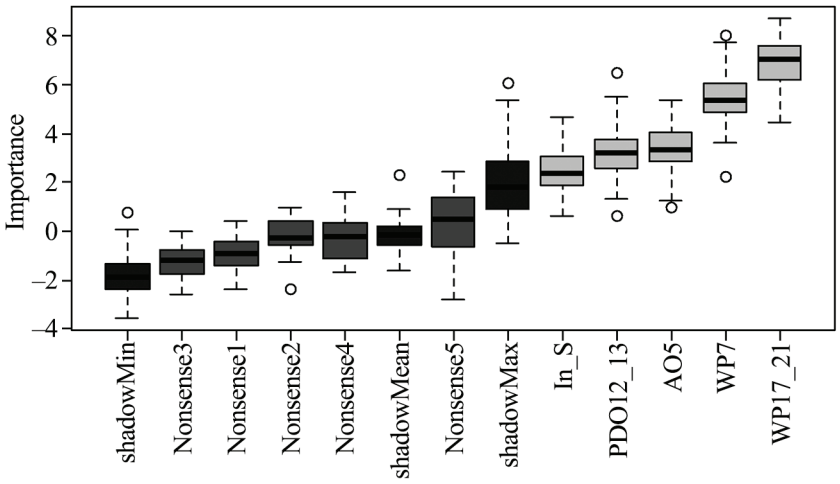

Рис. 10. Результат алгоритма Boruta: признаки, важность которых выше порога shadowMax признаются подтвержденными (светло-серый цвет), ниже порога — неподтвержденными (темно-серый цвет)
Fig. 10. Result of the algorithm of Boruta: the signs with importance higher the shadowMax threshold are reckoned as confirmed (light grey), and lower – unconfirmed (dark grey)

Естественно, что в лесе с композицией из пяти рассматриваемых предикторов ошибка в 4% обусловлена только уникальным сочетанием наблюдений в бутстреп-выборках и, соответственно, в тестовых наборах данных. Чтобы вычислить распределение ошибки OOB конкретно для случайных лесов с выбранными пятью признаками, мы провели расчет еще 1000 вариантов именно этой модели на различных зернах ГСЛ (скрипт 4). В качестве результатов для каждого леса отбирались ошибка OOB и прогноз уровня численности для поколения 2018 г. рождения (подход 2020 г.).

Распределение ошибки OOB в среднем составило 13,1% со стандартным отклонением 3,6%.

Скрипт 4

```
 ## создаем вектор для записи результатов ошибки OOB
vecOOB <- c()
 ## создаем дата-фрейм для записи результатов прогноза
DataPred <- data.frame(low = 1, average = 2, high = 3)

 ## НАЧАЛО ЦИКЛА
for(i in 1:1000) {
    ## подключаем библиотеку случайного леса
  library(randomForest)
    ## устанавливаем  зерно ГСЛ
  set.seed(i)
    ## создаем объект случайного леса
  resultData.rf <- randomForest(strata~ .,  xtest=NULL, ytest=NULL,
```

```
                        data=resultData, mtry=2, nodesize=1,
                        ntree=150,
                        replace=TRUE, importance=TRUE,
                        localImp=TRUE,
                        proximity=TRUE)
    ## записываем значение ошибки OOB в вектор
   vecOOB <- c(vecOOB, resultData.rf$err.rate[resultData.rf$ntree])
    ## вносим прогнозные данные
   predictData2018 <- data.frame(PDO12_13 = -0.27,
                        WP7 = -0.81,
                        WP17_21 = -0.83,
                        AO5 = 1.18,
                        ln_S = 3.31)
    ## вычисляем прогноз уровня на 2020 г.
   Pred <- predict(resultData.rf, predictData2018, type = “prob”)
    ## записываем результат прогноза в дата-фрейм
   DataPred <- rbind(DataPred, Pred)
  ## КОНЕЦ ЦИКЛА
}
  ## удаляем инициализирующую строку в дата-фрейме прогноза
DataPred <- DataPred[-1,]
  ## экспорт данных в MS Excel
library(xlsx)
write.xlsx(DataPred, file=”DataPred.xlsx”,
        sheetName=”Sample_Sheet”, row.names=F, showNA=F)
write.xlsx(vecOOB, file=”vecOOB.xlsx”,
        sheetName=”Sample_Sheet”, row.names=F, showNA=F)
```

Прогноз для уровня численности на 2020 г. для всех вариантов одинаковый, все леса проголосовали за высокий уровень численности high. Доля отданных голосов за уровень high в среднем по 1000 моделям случайных лесов составила 85,5% со стандартным отклонением 2,8%.

ЗАКЛЮЧЕНИЕ

Как показала данная работа, ансамблевые методы, в частности метод Random Forest, можно вполне успешно применять в задачах прогнозирования рыбных запасов. Средняя ошибка прогноза уровня численности в 13% относительно невелика и существенно ниже существующих практик прогнозирования горбуши, даже учитывая собственные дисперсии уровней воспроизводства стратифицированной модели на рис. 1. Оформление работы в виде программного кода на языке R значительно облегчит процесс и уменьшит время принятия решения. Тем не менее показанный прогноз горбуши северо-востока Камчатки на 2020 год не оправдался. Подход данной рыбы состоялся на низком уровне (low) — порядка 28 млн экз. При этом другие методы прогнозирования, такие как общая регрессионная модель, индикаторы ската в реперных реках, данные по траловой съемке молоди перед откочевкой в места зимнего нагула, говорили в совокупности о среднем уровне численности возврата данного поколения. Разброс таких прогнозов составил от 55 до 120 млн. Данное обстоятельство служит хорошим примером того, что не следует полагаться на какой-либо отдельный метод прогнозирования, но и учитывать также данные по другим методам.

Что послужило основанием для такой ошибки показанного метода — пока не ясно. Это может быть как и ошибкой метода, так и вмешательством неучтенного фактора. Понятно также, что используемые в качестве предикторов индексы достаточно масштабны и не могут отображать региональные климатические аномалии на сравнительно небольших территориях, которые занимает горбуша, начиная с периода нереста вплоть до ската и раннего морского периода жизни. Автор полагает, что включение в набор предикторов таких региональных показателей должно увеличить прогностические качества предложенного алгоритма.

ПРИЛОЖЕНИЕ

Основной скрипт

```
  ## подключаем библиотеку экспорта данных из таблицы Excel
library(readxl)
  ## экспортируем исходные данные в дата-фрейм
allData <- read_excel("~/R/randomForest/allData.xlsx")
  ## подключаем библиотеку dplyr, облегчает манипуляции с данными
library(dplyr)
  ## создаем дата-фреймы для записи результатов
DataVecOOB <- data.frame(iteration = 1:37)
DataPredictors <- data.frame(n = 1:5)

   ## НАЧАЛО ОСНОВНОГО ЦИКЛА
for(i in 1:1000) {
   ## преобразуем все символьные переменные в фактор,
   ## и объявляем новый дата-фрейм,равный исходному
  newData <- allData %>% mutate_if(is.character, as.factor)
   ## обозначаем порядок фактора (low < average < high)
  newData$strata <- ordered(newData$strata,levels = c("low", "average", "high"))
   ## создаем пустой вектор для записи промежуточных результатов
  vecOOB <- c()        ## для ошибки OOB каждого леса

   ## Во фрейме данных 69 объясняющих переменных.
   ## Чтобы уменьшить их количество будем выращивать лес,
   ## исключать самые незначимые предикторы (5%-квантиль по важности),
   ## и обучать снова.
   ## Цикл закончится, когда объясняющих переменных останется две

    ## НАЧАЛО ВЛОЖЕННОГО ЦИКЛА
   while (ncol(newData)-1 > 1) {

      ## формула для параметра mtry - квадратный корень из количества
      ## объясняющих переменных (округленный в большую сторону)
     X <- round(sqrt(ncol(newData)-1)) ## -1 вычесть зависимую переменную
     if (X < 2) X = 2 ## mtry не менее двух признаков
      ## формула для параметра nodesize - натур логарифм из количества
      ## объясняющих переменных минус единица (округленный в большую сторону)
     Y <- round(log(ncol(newData)-1)-1)
     if (Y < 1) Y = 1
      ## подключаем библиотеку случайного леса
     library(randomForest)
      ## зерно ГСЛ изменяется от 1 до 1000
     set.seed(i)
      ## создаем объект случайного леса
     newData.rf <- randomForest(strata~ .,  xtest=NULL, ytest=NULL,
                          data=newData, mtry=X, nodesize=Y,
                          ntree=150,
                          replace=TRUE,
                          importance=TRUE, localImp=TRUE,
                          proximity=TRUE)
      ## именуем данные по важности признаков,
     imp <- importance(newData.rf, type = 2) ## type = 2 - данные по 'mean decrease accuracy'
      ## устанавливаем критерий для фильтрации,
```

```
      ## находим 5%-квантиль, по которому будем отсекать незначимые предикторы
    quantileProc <- quantile(imp, 0.05)
      ## представляем imp как датафрейм
    impDataFrame <- as.data.frame(imp,row.names = NULL, optional = FALSE,
                          make.names = TRUE,
                          stringsAsFactors = default.stringsAsFactors())
      ## добавляем в него вектор индексов
    impDataFrame <- data.frame(MeanDecreaseGini = impDataFrame,
                         index = rownames(impDataFrame))
      ## сортируем по уменьшению важности
    impDataFrame <- impDataFrame[order(impDataFrame$MeanDecreaseGini, decreasing = TRUE),]
       ## если в дата-фрейме остается 5 предикторов, запоминаем их имена
      if ((ncol(newData)-1) == 5) {
       Predictors <- impDataFrame$index
      }
      ## фильтруем датафрейм важности по критерию quantileProc
    filterImp <- filter(impDataFrame, impDataFrame$MeanDecreaseGini <= quantileProc)
      ## представляем filterImp как вектор
    filterImpVector <- as.vector(filterImp$index)
      ## находим индексы совпадений
    I <-match(filterImpVector, names(newData), nomatch = 0, incomparables = NULL)
      ## записываем значение ошибки OOB в вектор
    vecOOB <- c(vecOOB, newData.rf$err.rate[newData.rf$ntree])
      ## удаляем признаки с индексом(-ами) I
    newData <- newData[ , -c(I)]

   ## КОНЕЦ ВЛОЖЕННОГО ЦИКЛА
  }
   ## добавляем промежуточные результаты в датафреймы результатов
  DataVecOOB <- cbind(DataVecOOB, vecOOB)
  DataPredictors <- cbind(DataPredictors, Predictors)

  ## КОНЕЦ ОСНОВНОГО ЦИКЛА
}
  ## удаляем инициализирующие векторы в дата-фреймах результатов
DataVecOOB <- DataVecOOB[, -1]
DataPredictors<- DataPredictors[, -1]
  ## Экспорт в MS excel
library(xlsx)
write.xlsx(DataVecOOB, file="OutputOOB.xlsx",
        sheetName="Sample_Sheet", row.names=F, showNA=F)
write.xlsx(DataPredictors, file="OutputPredictors.xlsx",
        sheetName="Sample_Sheet", row.names=F, showNA=F)
```

СПИСОК ЛИТЕРАТУРЫ

Бугаев А.В., Тепнин О.Б., Радченко В.И. 2018. Климатическая изменчивость и продуктивность тихоокеанских лососей Дальнего Востока России // Исслед. водн. биол. ресурсов Камчатки и сев.-зап. части Тихого океана. Вып. 49. С. 5–50.

Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р. 2016. Введение в статистическое обучение с примерами на языке R. Пер. С.Э. Мастицкого. М.: ДМК Пресс, 450 с.

Карпенко В.И. 1998. Ранний морской период жизни тихоокенских лососей: Монография. М.: ВНИРО, 165 с.

Кляшторин Л.Б., Любушин А.А. 2005. Циклические изменения климата и рыбопродуктивности. М.: ВНИРО, 235 с.

Маркевич Н.Б., Виленская Н.И. 1998. Влияние сроков нереста и термического режима на выживаемость и рост молоди горбуши *Oncorhinchus gorbusha* на ключевых и русловых нерестилищах Западной Камчатки // Исслед. биологии и динамики численности промысловых рыб Камчатского шельфа. Вып. I. Ч. 1. С. 85–104.

Радченко В. 2017. Открытый курс машинного обучения. Тема 5. Композиции: бэггинг, случайный лес. Электронный блог компании Open Data Science, адрес доступа: https://habr.com/ru/company/ods/blog/324402/.

Фельдман М.Г., Шевляков Е.А. 2015. Выживаемость камчатской горбуши как результат совокупного воздействия плотностной регуляции и внешних факторов среды // Изв. ТИНРО. Т. 182. С. 88–114.

Фельдман М.Г., Шевляков Е.А., Артюхина Н.Б. 2018. Оценка ориентиров пропуска производителей тихоокеанских лососей в бассейнах рек Северо-Восточной Камчатки // Исслед. водн. биол. ресурсов Камчатки и сев.-зап. части Тихого океана. Вып. 51. С. 5–26.

Шитиков В.К., Мастицкий С.Э. 2017. Классификация, регрессия, алгоритмы Data Mining с использованием R. Электронная книга, адрес доступа: https://github.com/ranalytics/data-mining.

Шунтов В.П., Темных О.С. 2005. Основные результаты изучения морского периода жизни тихоокеанских лососей в ТИНРО-Центре // Изв. ТИНРО. Т. 141. С. 30–55.

Шунтов В.П., Темных О.С. 2011. Тихоокеанские лососи в морских и океанических экосистемах: Монография. Т. 2. Владивосток: ТИНРО-Центр, 473 с.

Шуровьески Дж. 2007. Мудрость толпы. Почему вместе мы умнее, чем поодиночке, и как коллективный разум формирует бизнес, экономику, общество и государство. Пер. с англ. М.: ООО «И.Д. Вильяме», 304 с.

Breiman L. 1996a. Bagging Predictors // Machine Learning: journal. Vol. 24, no. 2. P. 123–140.

Breiman L. 1996b. Out-of-bag estimation. Technical report, Dept. of Statistics, Univ. of Calif., Berkeley. Электронный источник, адрес доступа: https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf.

Breiman L. 2001. Random Forests // Machine Learning: journal. Vol. 45, no. 1. P. 5–32.

Breiman L., Friedman J.H., Olshen R.A., Stone C.J. 1984. Classification and regression trees. Wadsworth International Group, Belmont CA, 368 p.

Delgado F.M., Cernadas E., Barro S., Amorim D. 2014. Do we need hundreds of classifiers to solve real world classification problems? // J. of Machine Learning Research, № 15. P. 3133–3181.

Efron B. 1979. Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics, Vol. 7. № 1. P. 1–26.

Galton F. 1907. Vox populi // Nature, № 75. P. 450–451.

Haeseker S., Dorner B., Peterman R., Su Z. 2007. An improved sibling model for forecasting Chum Salmon and Sockeye Salmon abundance // North American Journal of Fisheries Management. № 27. P. 634–642.

Hare S.R. 1996. Low frequency climate variability and salmon production. Ph.D. Dissertation. Univ. of Washington, Seattle, WA, 306 p.

Hare S.R., Francis R.C. 1995. Climate change and salmon production in the Northeast Pacific Ocean // In Climate Change and Northern Fish Populations, ed. by R.J. Beamish, Can. Spec. Publ. Fish. Aquat. Sci. Vol. 121. P. 357–372.

Ho T.K. 1995. Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition. Montreal, QC: 14–16 August. P. 278–282.

Kleinberg E. 1990. Stochastic Discrimination // Annals of Mathematics and Artificial Intelligence, Vol. 1 (1–4). P. 207–239.

Kleinberg E. 1996. An Overtraining-Resistant Stochastic Modeling Method for Pattern Recognition // Annals of Statistics. Vol. 24 (6). P. 2319–2349.

Kursa M. 2020. Boruta for those in a hurry. Электронная статья, адрес доступа: https://cran.r-project.org/web/packages/Boruta/vignettes/inahurry.pdf.

Kursa M., Rudnicki W. 2010. Feature Selection with the Boruta Package // J. of Statistical Software. Vol. 36 (11). P. 2–12.

Linkin M.E., Nigam S. 2008. The North Pacific Oscillation – West Pacific Teleconnection Pattern: Mature-Phase Structure and Winter Impacts // J. Climate. Vol. 21. № 9. P. 1979–1997.

Mantua N., Hare S., Zhang Y., Wallace J., Francis R. 1997. A Pacific interdecadal climate oscillation with impacts on salmon production // Bull. Amer. Meteor. Soc., № 78. P. 1069–1079.

Mantua N.J., Hare S.R. 2002. The Pacific Decadal Oscillation // J. of Oceanography. Vol. 58. P. 35–44.

Paluszyńska A. Understanding random forests with randomForestExplainer. Электронная статья, адрес доступа: https://cran.rstudio.com/

web/packages/randomForestExplainer/vignettes/randomForestExplainer.html.

Peterman R.M. 1982. Model of salmon age structure and its use in preseason forecasting and studies of marine survival // Canadian Journal of Fisheries and Aquatic Sciences. № 39. P. 1444–1452.

Quinlan J.R. 1986. Induction of Decision Trees // Machine Learning. Kluwer Academic Publishers. № 1. P. 81–106.

Ricker W.E. 1954. Stock and Recruitment // J. of the Fisheries Research Board of Canada. Vol. 11. № 5. P. 559–623.

Thompson D., Wallace J. 1998. The Arctic Oscillation signature in the wintertime geopotential height and temperature fields. Geophys. Res. Lett., Vol. 25. № 9. P. 1297–1300.

REFERENCES

Bugaev A.V., Tepnin O.B., Radchenko V.I. Climate variability and Pacific salmon productivity in Russian Far East. *The researches of the aquatic biological resources of Kamchatka and of the north–west part of the Pacific Ocean*, 2018, vol. 49, pp. 5–50. (In Russian with English abstract)

James G., Whitton D., Hastie T., Tibshirani R. An Introduction to Statistical Learning with Applications. Moscow: DMK Press, 2016, 450 p.

Karpenko V.I. *Ranniy morskoy period zhizni tikhookenskikh lososey* [Early marine period of life of Pacific Ocean salmon: monograph]. Moscow: VNIRO, 1998, 165 p.

Klyashtorin L.B., Lyubushin A.A. *Tsiklicheskiye izmeneniya klimata i ryboproduktivnosti* [Cyclical changes in climate and fish productivity]. Moscow: VNIRO, 2005, 235 p.

Markevich N.B., Vilenskaya N.I. Effects of the time and thermal regime of spawning on the survival and growth of juvenile pink salmon *Oncorhynchus gorbusha* in the spring and river main body spawning grounds of Western Kamchatka. *The researches of the aquatic biological resources of Kamchatka and of the north-west part of the Pacific Ocean*, 2018, vol. 49, pp. 5–50. (In Russian)

Radchenko V. Machine Learning Open Course. Topic 5. Compositions: bagging, random forest. 2017, Electronic blog of the Open Data Science company, access address: https://habr.com/ru/company/ods/blog/324402/

Feldman M.G., Shevlyakov E.A. Survival of Kamchatka pink salmon as a result of combined effect of density and environmental regulating factors. *Izvestia TINRO*, 2015, vol. 182, pp. 88–114. (In Russian)

Feldman M.G., Shevlyakov E.A., Artukhina N.B. Evaluation of teh Pacific salmon spawning escapement parameters for the river basins of North-East Kamchatka. *The researches of the aquatic biological resources of Kamchatka and of the north-west part of the Pacific Ocean*, 2018, vol. 51, pp. 5–26. (In Russian)

Shitikov V.K., Mastitsky S.E. Classification, regression and R-using algorithms of data mining. Electronic book, 2017, access address: https://github.com/ranalytics/data-mining

Shuntov V.P., Temnykh O.S. Main results of the TINRO-Center research of marine period of Pacific salmon. *Izvestia TINRO*, 2005, vol. 141, pp. 30–55. (In Russian)

Shuntov V.P., Temnykh O.S. Pacific salmon in marine and ocean ecosystems: monograph. *Vladivostok: TINRO-Center*, 2011, vol. 2, 473 p.

Surowiecki J. The Wisdom of Crowds: Why the Many are Smarter than the the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. *New York: Anchor Books*, 2005. 304 p. (Translated into Russian)

Breiman L. Bagging Predictors. *Machine Learning*, 1996a, № 24, pp. 123–140.

Breiman L. Out-of-bag estimation. Technical report, Dept. of Statistics, Univ. of Calif., Berkeley, 1996b. Electronic source, access address: https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf.

Breiman L., Friedman J.H., Olshen R.A., Stone C.J. Classification and regression trees. *Wadsworth International Group, Belmont CA*, 1984, 368 p.

Delgado F.M., Cernadas E., Barro S., Amorim D. Do we need hundreds of classifiers to solve real world classification problems? *J. of Machine Learning Research*, 2014, № 15, pp. 3133–3181.

Efron B. Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics*, 1979, vol. 7. № 1, pp. 1–26.

Galton F. Vox populi. *Nature*, 1907, № 75, pp. 450–451.

Haeseker S., Dorner B., Peterman R., Su Z. An improved sibling model for forecasting Chum Salmon and Sockeye Salmon abundance. *North American Journal of Fisheries Management*, 2007, № 27, pp. 634–642.

Hare S.R. Low frequency climate variability and salmon production. *Ph.D. Dissertation. Univ. of Washington, Seattle, WA*, 1996, 306 p.

Hare S.R., Francis R.C. In Climate Change and Northern Fish Populations. *Can. Spec. Publ. Fish. Aquat. Sci.*, 1995, vol. 121, pp. 357–372.

Ho T.K. Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition. *Montreal, QC*, *14–16 August*, 1995, pp. 278–282.
Kleinberg E. Stochastic Discrimination. *Annals of Mathematics and Artificial Intelligence*, 1990, vol. 1 (1–4), pp. 207–239.
Kleinberg E. An Overtraining-Resistant Stochastic Modeling Method for Pattern Recognition. *Annals of Statistics*, 1996, vol. 24 (6), pp. 2319–2349.
Kursa M. Boruta for those in a hurry. 2020. Electronic source, access address: https://cran.r-project.org/web/packages/Boruta/vignettes/inahurry.pdf.
Kursa M., Rudnicki W. Feature Selection with the Boruta Package. *J. of Statistical Software*, 2010, vol. 36 (11), pp. 2–12.
Linkin M.E., Nigam S. The North Pacific Oscillation – West Pacific Teleconnection Pattern: Mature-Phase Structure and Winter Impacts. *J. Climate*, 2008, vol. 21, № 9, pp. 1979–1997.
Mantua N., Hare S., Zhang Y., Wallace J., Francis R.A. Pacific interdecadal climate oscillation with impacts on salmon production. *Bull. Amer. Meteor. Soc.*, 1997, № 78, pp. 1069–1079.
Mantua N.J., Hare S.R. The Pacific Decadal Oscillation. *J. of Oceanography*, 2002, vol. 58, pp. 35–44.
Paluszyńska A. Understanding random forests with randomForestExplainer. 2021. Electronic source, access address: https://cran.rstudio.com/web/packages/randomForestExplainer/vignettes/randomForestExplainer.html.
Peterman R.M. Model of salmon age structure and its use in preseason forecasting and studies of marine survival. *Canadian Journal of Fisheries and Aquatic Sciences*, 1982, № 39, pp. 1444–1452.
Quinlan J.R. Induction of Decision Trees. *Machine Learning. Kluwer Academic Publishers*, 1986, № 1, pp. 81–106.
Ricker W.E. Stock and Recruitment. *J. of the Fisheries Research Board of Canada*, 1954, vol. 11, № 5, pp. 559–623.
Thompson D., Wallace J. The Arctic Oscillation signature in the wintertime geopotential height and temperature fields. *Geophys. Res. Lett.*, 1998, vol. 25, № 9, pp. 1297–1300.

Статья поступила в редакцию: 22.10.2020
Статья принята после рецензии: 21.11.2020