



Московский государственный университет имени М.В. Ломоносова

Факультет космических исследований

Магистерская программа «Методы и технологии дистанционного
зондирования Земли»

КУРСОВАЯ РАБОТА

на тему: «Развитие методов классификации и оценки
значимости признаков на основе случайных лесов в контексте
задачи картографирования земного покрова России»

Абдуллаева Евгения Гасановна

Научный руководитель:
к.т.н. Хвостиков Сергей Антонович

Москва, 2022 г.

Аннотация

Abstract

Содержание

1 Введение	5
2 Анализ данных	6
2.1 Источник данных	6
2.2 Общие сведения о данных	7
2.3 Классифицируемые типы земного покрова	8
2.3.1 Тематические классы земного покрова и их пред- ставленность в выборке	8
2.3.2 Визуализация тематических классов земного покрова	12
2.4 Инструменты для анализа данных	34
3 Классификация земного покрова с помощью случайного леса	35
3.1 Математическая постановка задачи классификации	35
3.2 Численная оценка качества алгоритма	36
3.2.1 Accuracy	37
3.2.2 Precision, Recall	37
3.2.3 F-score	38
3.2.4 Confusion Matrix	39
3.2.5 Receiver Operating Characteristic Curve, Area Under Receiver Operating Characteristic Curve	40
3.2.6 Out-of-Bag Error	41
3.3 Случайный лес	41
3.3.1 Модель случайного леса	42
3.3.2 Предпосылки использования случайного леса	43
3.3.3 Обучение случайного леса и оценка качества клас- сификации	43
3.3.4 Оптимизация гиперпараметров алгоритма	45

3.3.5 Проверка случайного леса на полной тестовой выборке	46
4 Оценка значимости признаков	49
4.1 Оценка значимости признаков на основе информации, предоставляемой моделью случайного леса	49
4.2 Оценка значимости признаков с помощью теста χ^2	50
4.3 Оценка значимости признаков с помощью рекурсивного исключения	51
4.4 Обучение случайного леса и оценка качества классификации на выборке, содержащей наиболее значимые признаки	52
5 Заключение	55
6 Литература	56
7 Приложения	57

1 Введение

2 Анализ данных

2.1 Источник данных

В данной работе классификация осуществляется на основе спутниковых данных (стандартного продукта MOD09) 2010 года (весенних, летних и осенних) и 2011 года (зимних), полученных спектрорадиометром Moderate Resolution Imaging Spectroradiometer (MODIS), установленным на спутниках Terra и Aqua. Продукт MOD09 представляет собой данные о спектрально-отражательных характеристиках земной поверхности, для которых была выполнена геопривязка и корректировка на атмосферу. Данные предварительно обработаны для исключения влияния облаков и теней от них путем осреднения значений яркости за сезон, по ним созданы сезонные композиты.

Цитируя монографию [1], укажем, что прибор MODIS разработан для изучения биологических и физических процессов в глобальном масштабе с периодичностью наблюдений в 1-2 дня, в частности, для исследований растительного покрова. MODIS имеет 36 спектральных каналов в диапазоне $\lambda = 0.46\text{-}14.39$ мкм, в том числе информативные для изучения растительности красный ($\lambda = 0.62\text{-}0.67$ мкм) и ближний инфракрасный ($\lambda = 0.84\text{-}0.88$ мкм) каналы с пространственным разрешением 250 м, и ряд каналов с разрешением 500 м, используемых для анализа характеристик растительности и фильтрации облачности. Полоса охвата прибора составляет 2330 км, а покрытие данными измерений всей территории России обеспечивается с периодичностью не реже одного раза в сутки. Таким образом, данные прибора образуют непрерывный однородный архив ежедневных наблюдений в течение более 15 лет, анализ которых может быть эффективно использован для изучения и мониторинга растительного покрова. Данные MODIS успешно используются для создания глобальной карты типов земного покрова.

2.2 Общие сведения о данных

Данные для классификации представлены в табличном виде. Полный набор данных содержит 74029669 элементов (строк таблицы). Каждый элемент описан 14-ю признаками (столбцы таблицы), содержащими неотрицательные целочисленные значения:

CLASS (значения от 1 до 23) — индекс класса элемента выборки;

X (значения от 5820 до 40161), Y (значения от 1429 до 20580) — координаты элемента выборки (индекс пикселя в растровом изображении, размер пикселя — 230 метров);

WINTER1 (значения от 0 до 10583), WINTER2 (значения от 0 до 10584) — яркость композитного изображения за зимний сезон в красном канале (RED, $\lambda = 0.62\text{-}0.67$ мкм) и ближнем инфракрасном канале (NIR, $\lambda = 0.84\text{-}0.88$ мкм), соответственно;

SPRING1 (значения от 1 до 8653) — яркость композитного изображения за весенний сезон в канале RED;

SPRING2 (значения от 1 до 7018) — яркость композитного изображения за весенний сезон в канале NIR;

SPRING3 (значения от 1 до 5211) — яркость композитного изображения за весенний сезон в коротковолновом инфракрасном канале (SWIR, $\lambda = 1.63\text{-}1.65$ мкм);

SUMMER1 (значения от 1 до 8653), SUMMER2 (значения от 1 до 6907), SUMMER3 (значения от 1 до 4923) — яркость композитного изображения за летний сезон в каналах RED, NIR, SWIR, соответственно;

FALL1 (значения от 1 до 8653), FALL2 (значения от 1 до 6907), FALL3 (значения от 1 до 5280) — яркость композитного изображения за осенний сезон в каналах RED, NIR, SWIR, соответственно.

Классификация выборки изначально проводилась с помощью алгоритма Locally Adaptive Global Mapping Algorithm (LAGMA), формально описанного в [1], методом максимального правдоподобия. Результаты автоматической классификации были подвергнуты эксперту визуаль-

ному анализу.

2.3 Классифицируемые типы земного покрова

2.3.1 Тематические классы земного покрова и их представлена- ность в выборке

Представленные в наборе данных типы земного покрова включают в себя 19 тематических классов, образующих 5 различных групп земно-го покрова, описание групп и классов приведено ниже в соответствии с монографией [1]:

1. Леса:

- Темнохвойные вечнозеленые насаждения (*темнохвойный лес*), в пологе которых не менее 80% площади крон состав-ляют теневыносливые виды хвойных деревьев, включая ель, пихту и сибирскую сосну (кедр).
- Светлохвойные вечнозеленые насаждения (*светлохвойный лес*), в пологе которых не менее 80% площади крон состав-ляют деревья сосны обыкновенной.
- Лиственные насаждения (*лиственный лес*), в пологе которых не менее 80% площади занимают кроны березы и осины, а также широколиственных пород, включая дуб, липу, ясень, клен, вяз и некоторые другие виды.
- Смешанные насаждения с преобладанием хвойных пород (*смешанный лес с преобладанием хвойных*), в которых кроны хвойных деревьев занимают от 60% до 80%, а лиственных — от 20% до 40% площади полога.
- Смешанные насаждения (*смешанный лес*), в которых площа-ди крон хвойных и лиственных пород деревьев представлены примерно в равных пропорциях (40-60%) в пологе.

- Смешанные насаждения с преобладанием лиственных пород (*смешанный лес с преобладанием лиственных*), в которых кроны лиственных пород деревьев занимают от 60% до 80%, а хвойных — от 20% до 40% площади полога.
- Хвойные листопадные (лиственничные) насаждения (*хвойный листопадный лес*), в пологе которых кроны деревьев лиственницы занимают более 80% площади.
- *Редины хвойные листопадные* (лиственничные), представляющие собой участки, занятые отдельно стоящими деревьями или разреженными насаждениями лиственницы с проектным покрытием крон менее 20%.

2. Травяно-кустарниковая растительность:

- *Луга* — травяная растительность с продолжительностью вегетационного сезона более 5 месяцев, видовой состав которой характеризуется господством многолетних трав, главным образом злаков и осоковых, в условиях достаточного увлажнения. Площадь проекции крон деревьев и кустарников на земную поверхность составляет менее 20%.
- *Степь* — травяной покров образован преимущественно засухоустойчивыми многолетними дерновинными злаками (ковыль, типчак, полынь, житняк и др.). Встречается большое разнообразие видов степных кустарников и полукустарников, а также короткоцветущих эфемероидов и эфемеров.
- Хвойные вечнозеленые кустарники (*хвойный кустарник*) — кустарниковые заросли или низкоствольные леса из кедрового стланика.
- Лиственные кустарники (*лиственный кустарник*) — сообщество низкорослых и стелющихся кустарников (кустарниковых или карликовых берез, полярных ив и др.).

3. Тундра:

- *Кустарничковая тундра* — сухая тундра с редкой фрагментарной растительностью, среди которой доминируют виды альпоарктических кустарничковых сообществ высотой менее 15 см. Распространены также мохово-лишайниковый покров и разнотравье.
- *Травянистая тундра* представлена главным образом различными видами трав и мхов, произрастающими на сырых почвах и образующими сплошной растительный покров. Часто встречаются кустарнички высотой до 40 см.
- *Кустарниковая тундра* с доминированием кустарников (карликовая береза и различные виды ивы) высотой более 40 см, иногда с примесью можжевельника, ольхи или кедрового стланника.

4. Водно-болотные комплексы:

- *Болота* — территории, характеризующиеся избыточным увлажнением с преобладанием растительного покрова из мхов, лишайников, тростника, осоки и некоторых других видов. Часто встречаются участки с наличием редкого (<20%) древесного полога.
- *Прибрежная растительность* — гидрофильная травяная и древеснокустарниковая растительность по берегам водоемов, часто периодически затопляемая.

5. Не покрытые растительностью земли:

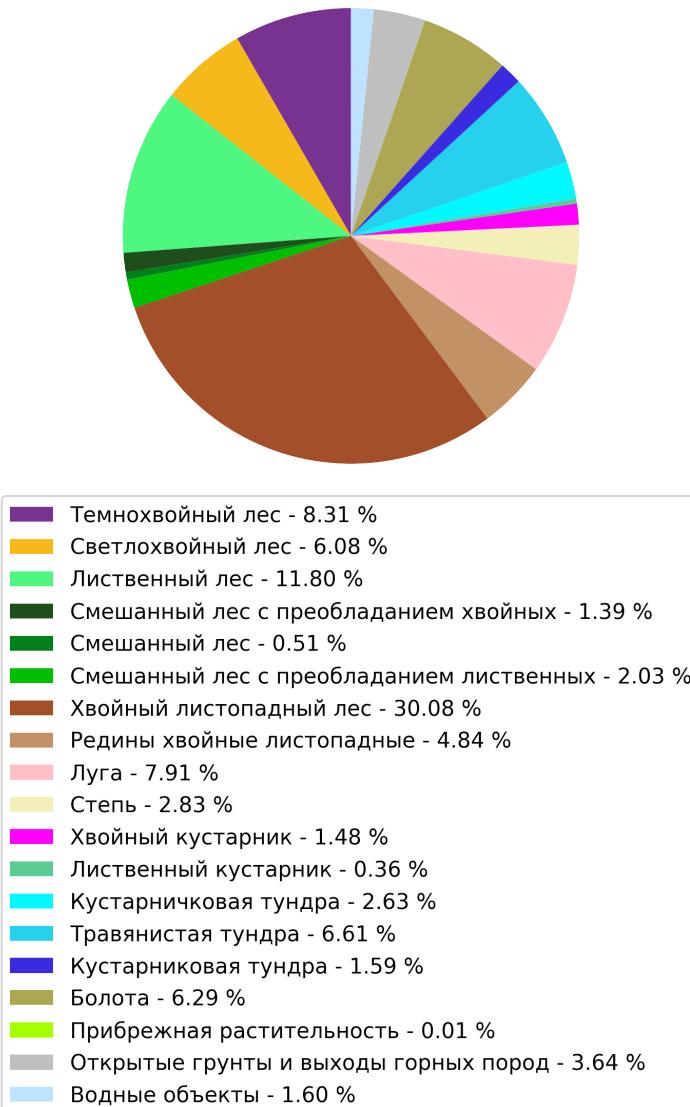
- *Открытые грунты и выходы горных пород* — земли, суммарное проективное покрытие которых растительностью всех видов не превышает 20%.

- *Водные объекты* — речные и озерные внутренние водоемы, а также прибрежные участки открытой воды.

Таблица 1: Количество представителей тематических классов земного покрова в выборке

Класс	Количество
Темнохвойный лес	6153034
Светлохвойный лес	4500884
Лиственный лес	8734036
Смешанный лес с преобладанием хвойных	1025492
Смешанный лес	378946
Смешанный лес с преобладанием лиственных	1505855
Хвойный листопадный лес	22271771
Редины хвойные листопадные	3586296
Луга	5858059
Степь	2091630
Хвойный кустарник	1096112
Лиственный кустарник	267236
Кустарниковая тундра	1943579
Травянистая тундра	4893560
Кустарниковая тундра	1178381
Болота	4657062
Прибрежная растительность	9661
Открытые грунты и выходы горных пород	2691471
Водные объекты	1186604
Всего элементов	74029669

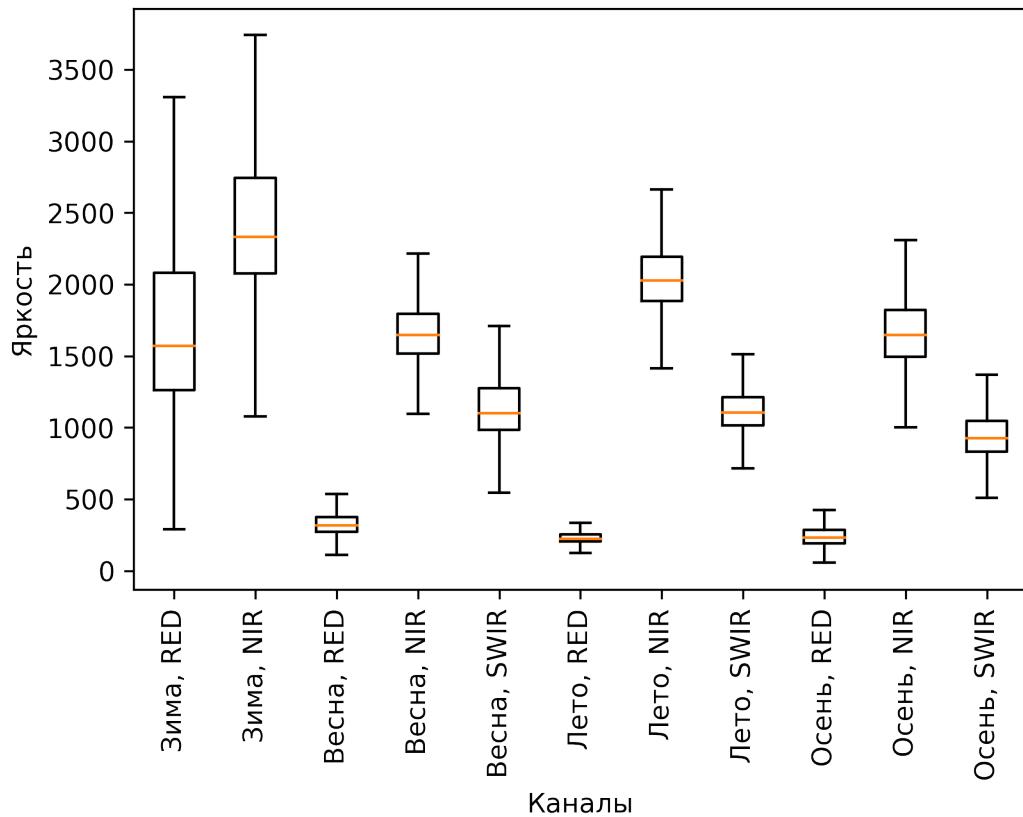
Рис. 1: Относительное количество представителей тематических классов земного покрова в выборке



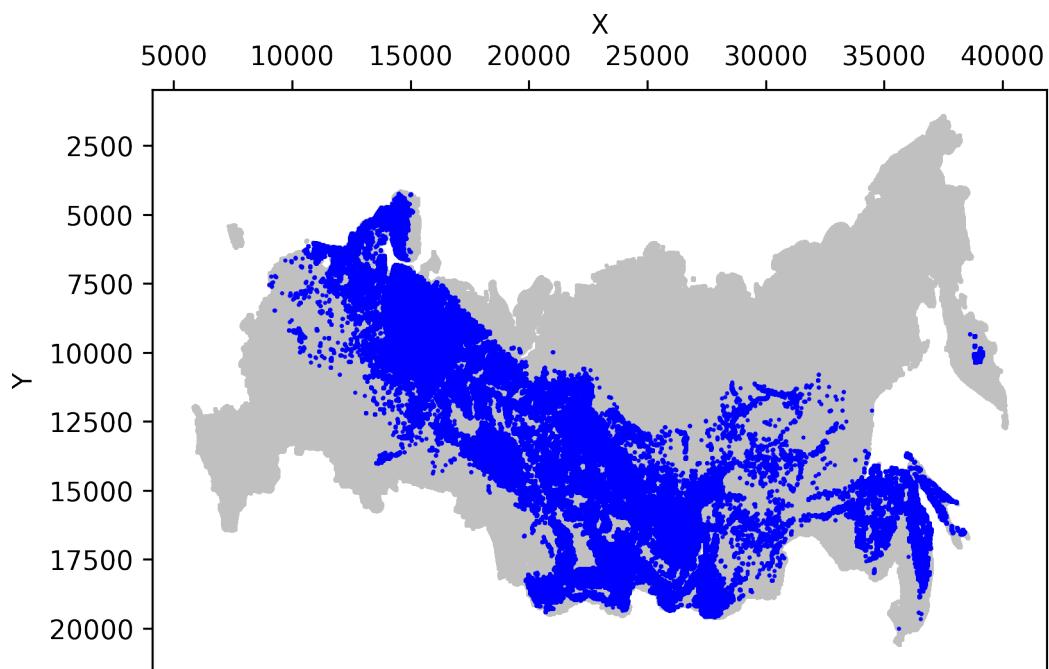
2.3.2 Визуализация тематических классов земного покрова

Рис. 2: Визуализация тематических классов земного покрова, представленных в выборке. Для каждого класса на первом графике представлена диаграмма размаха, показывающая медиану, нижний и верхний квантили, минимальное и максимальное значение сезонной яркости в различных спектральных каналах; на втором графике представлена карта пространственного распределения элементов-представителей класса (синим).

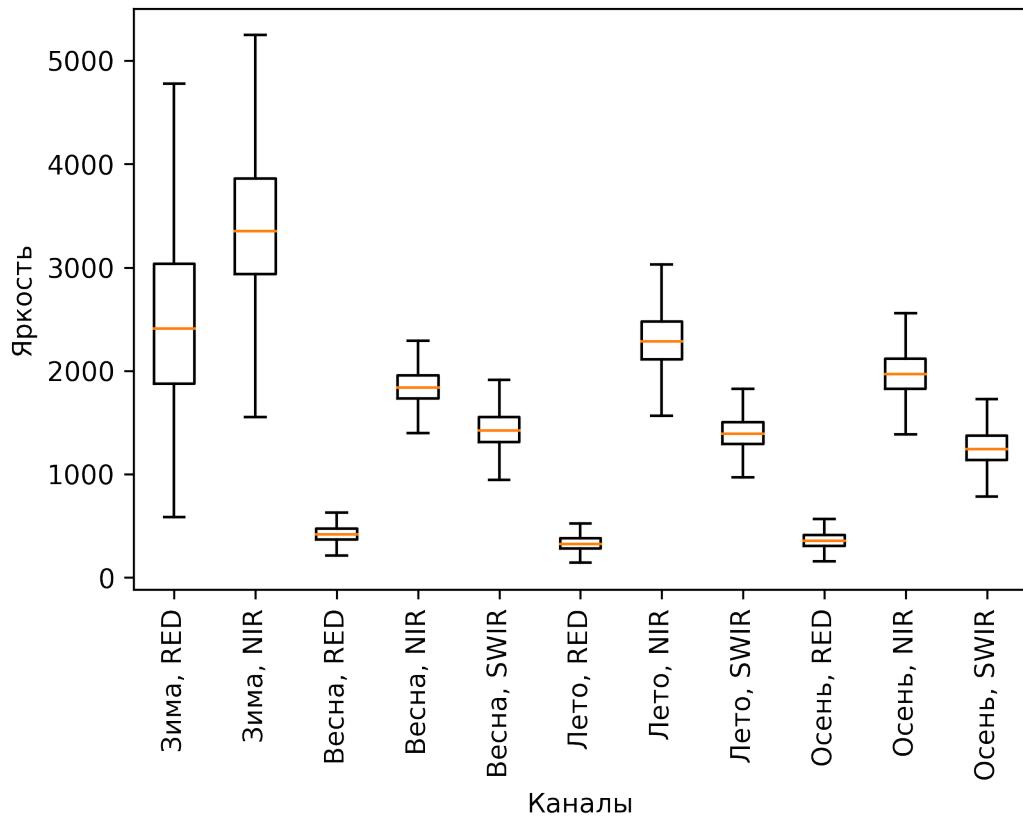
Темнохвойный лес



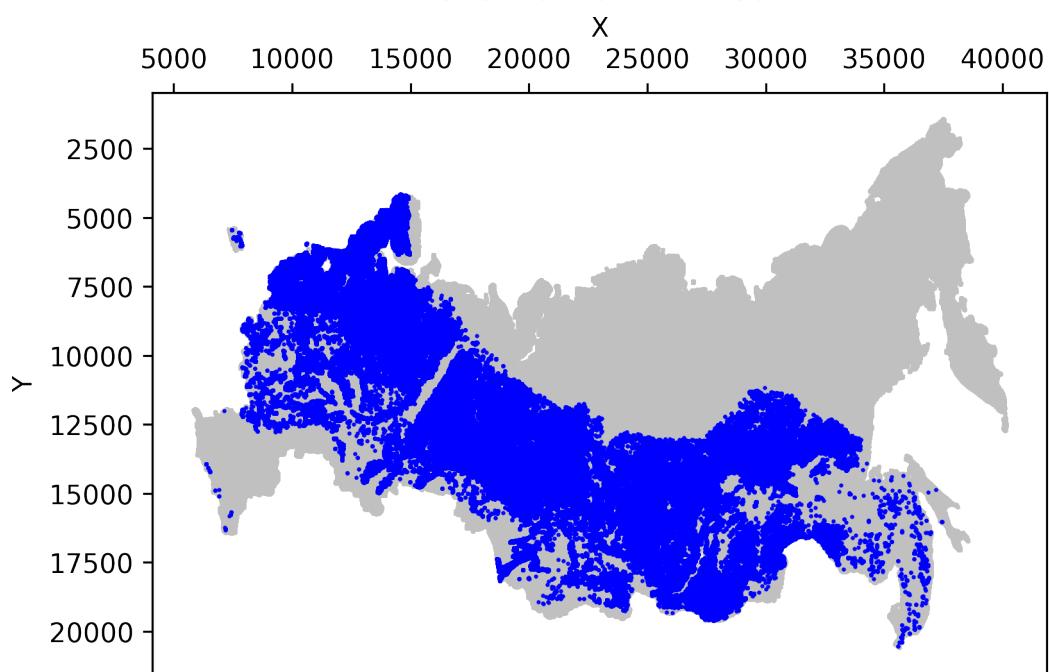
Темнохвойный лес



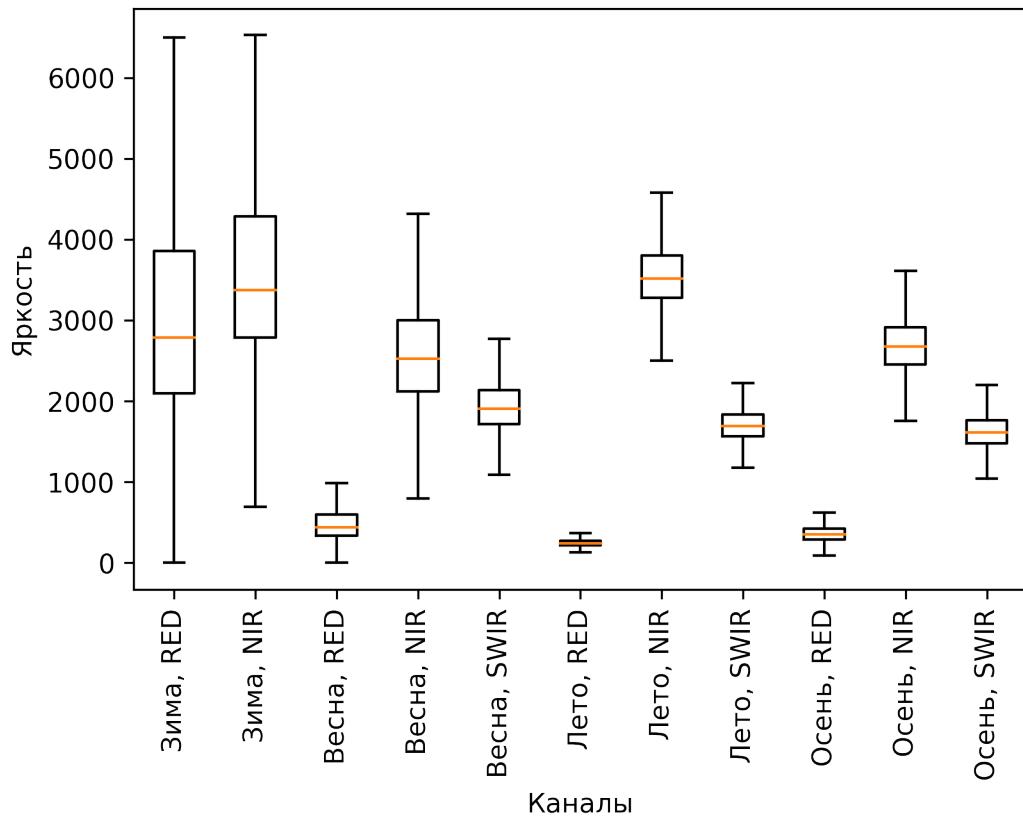
Светлохвойный лес



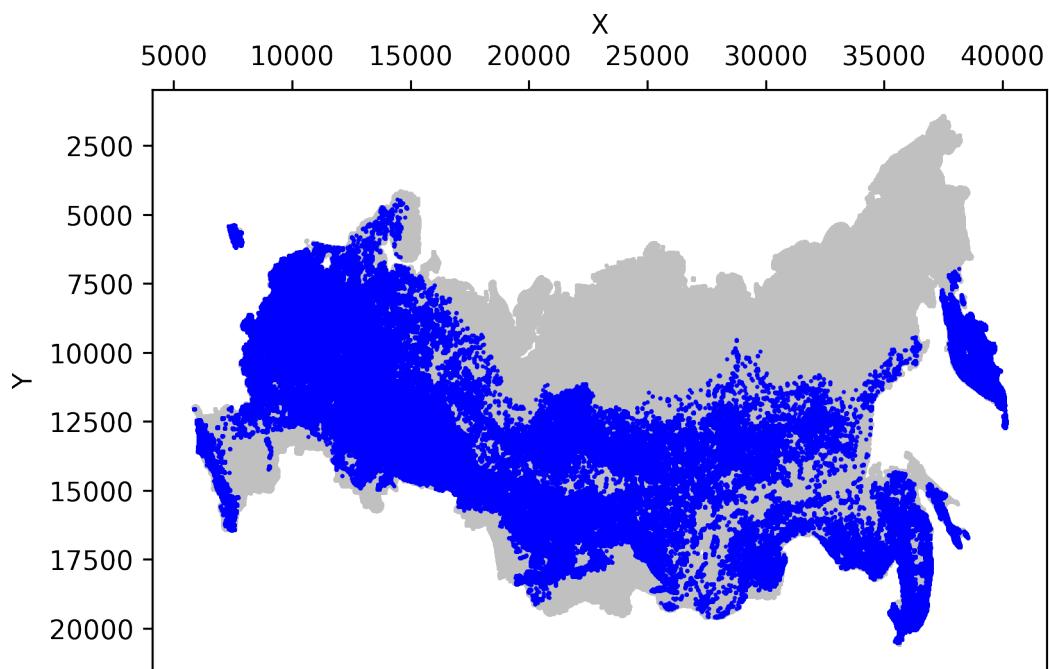
Светлохвойный лес



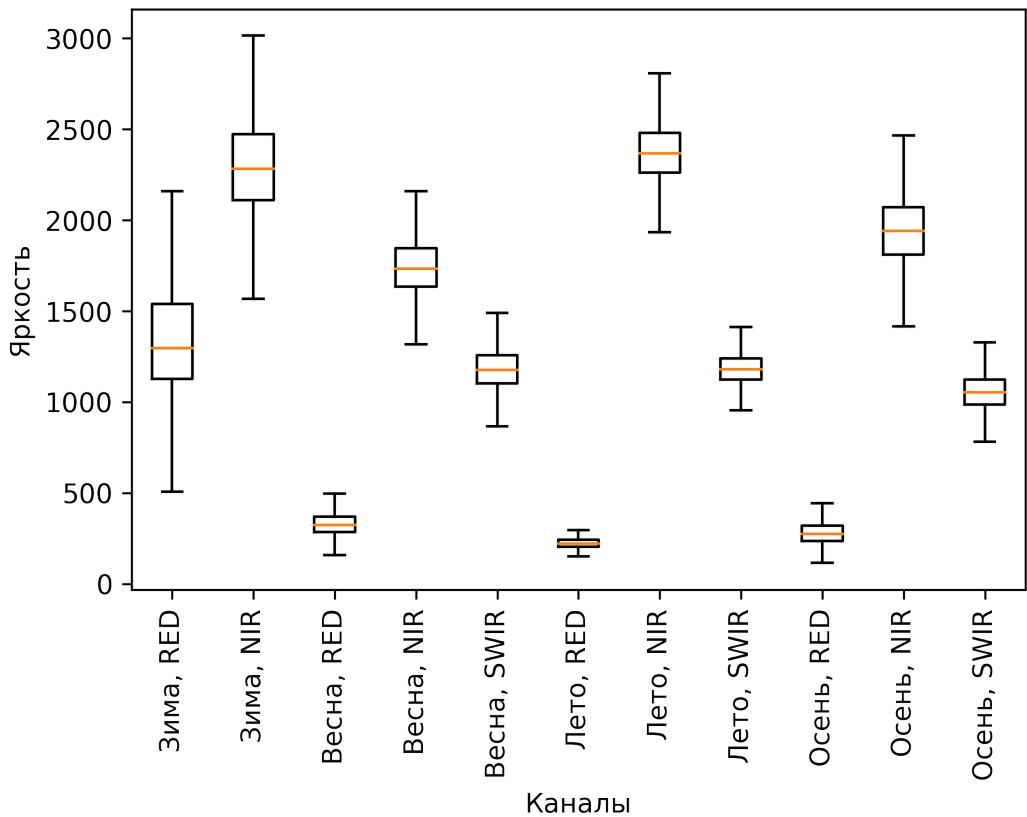
Лиственый лес



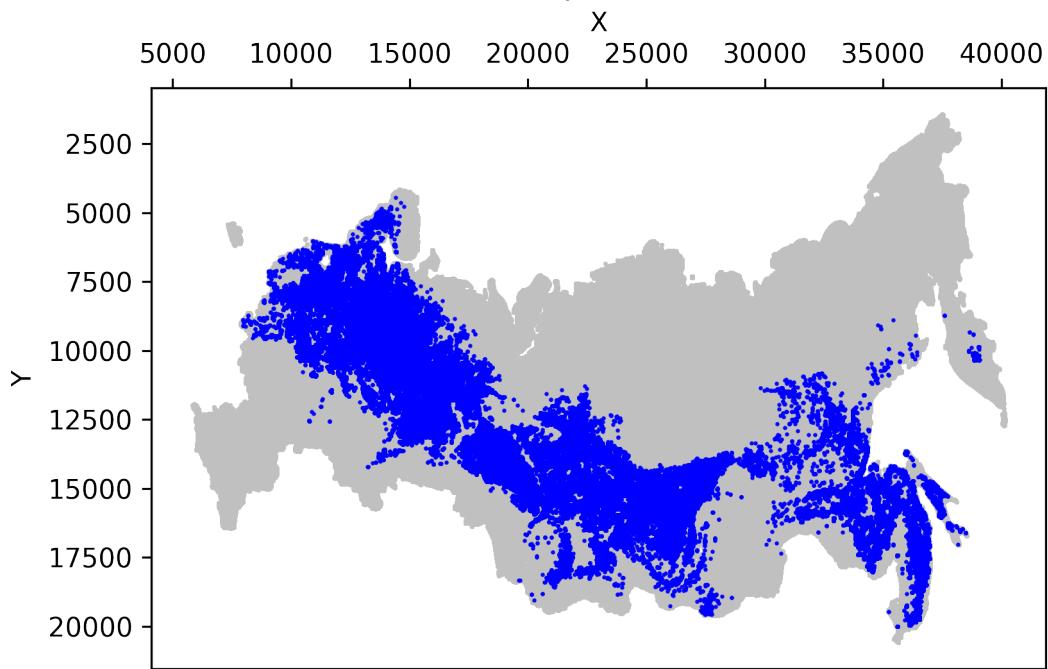
Лиственый лес



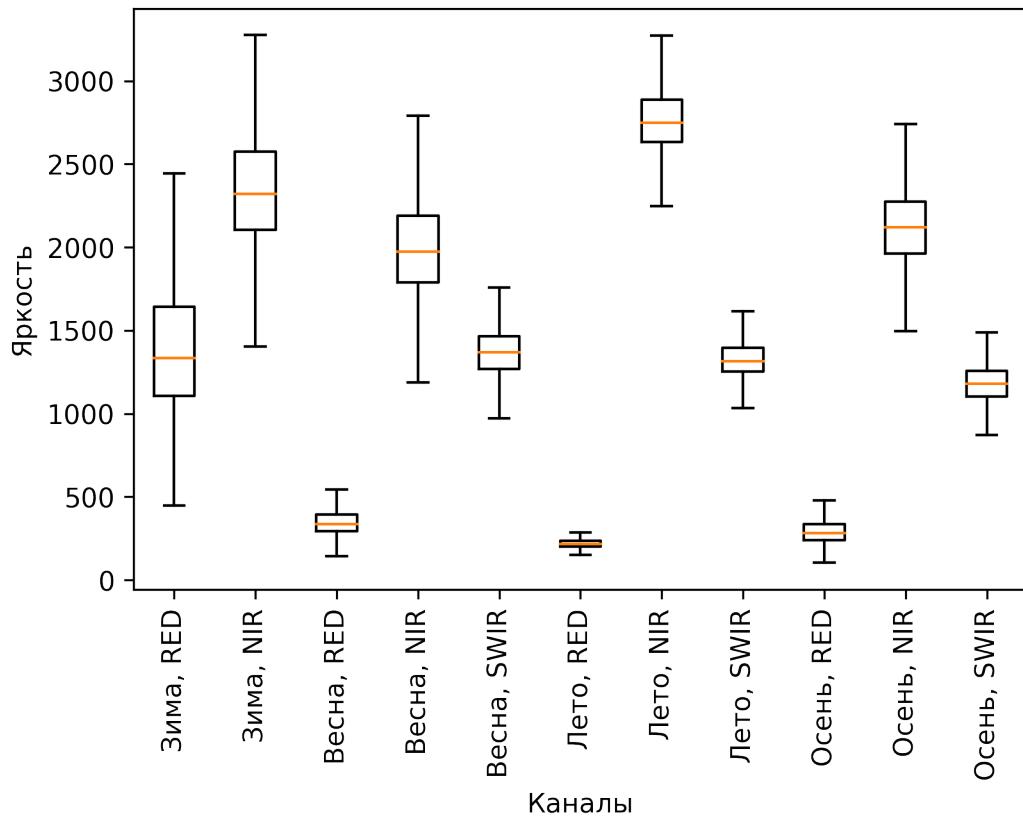
Смешанный лес с преобладанием хвойных



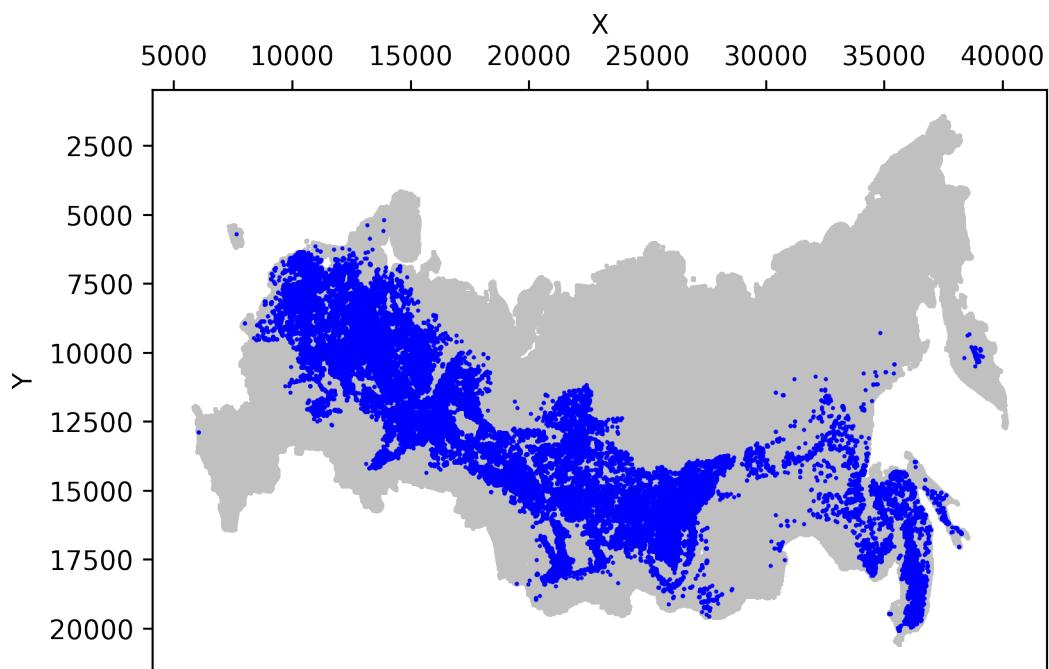
Смешанный лес с преобладанием хвойных



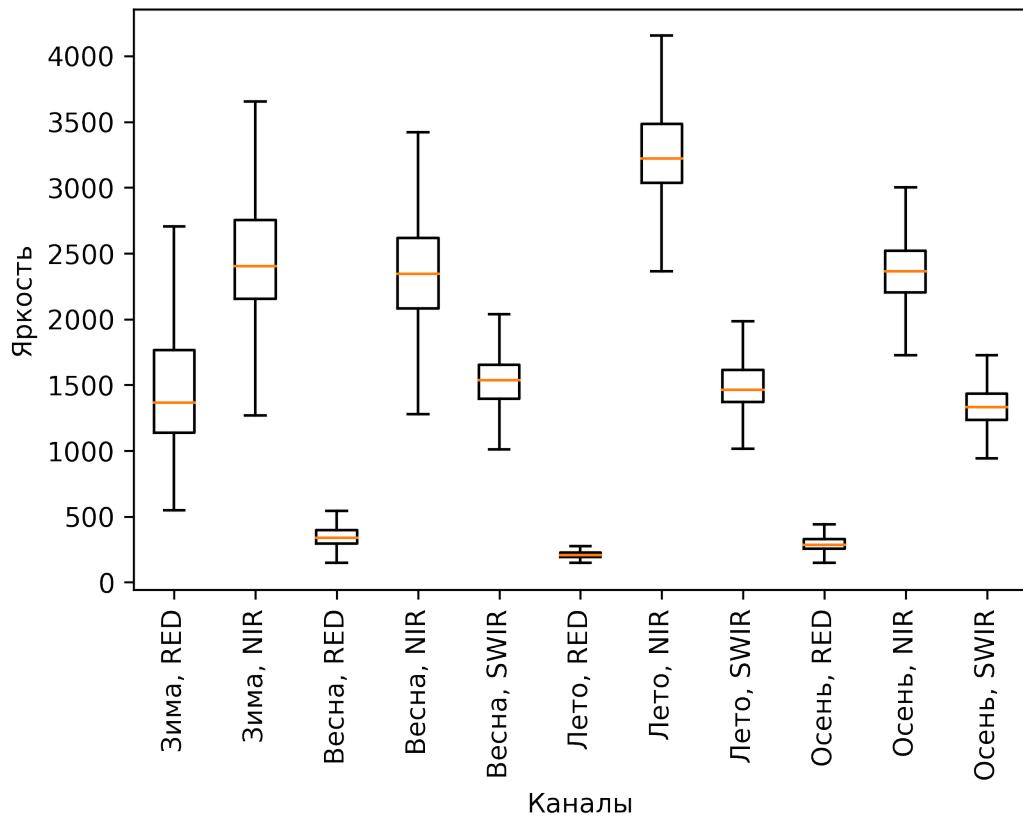
Смешанный лес



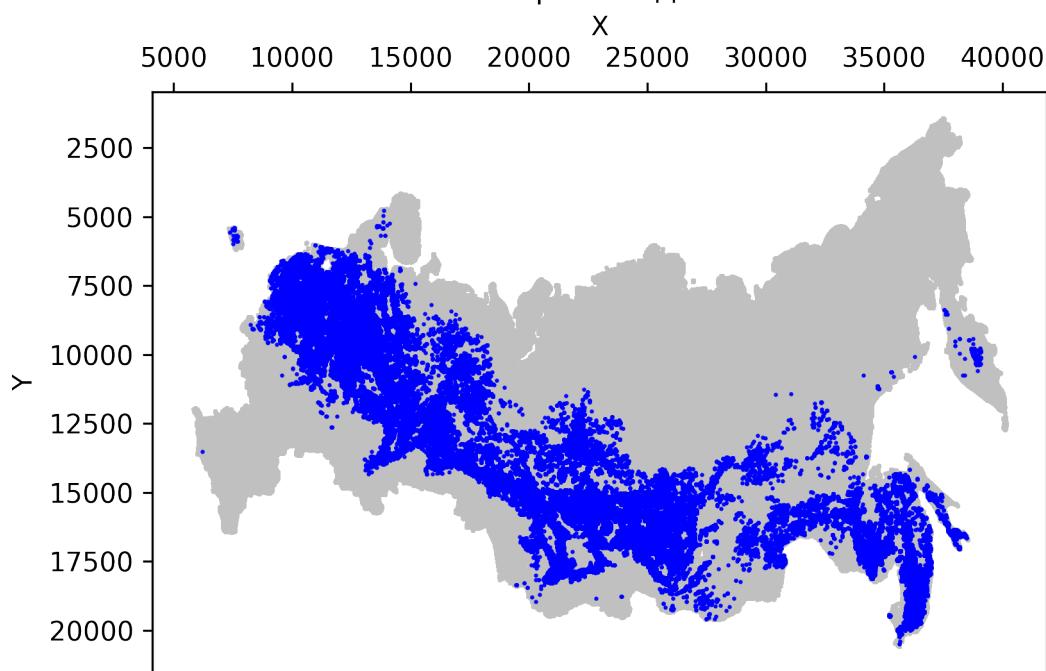
Смешанный лес



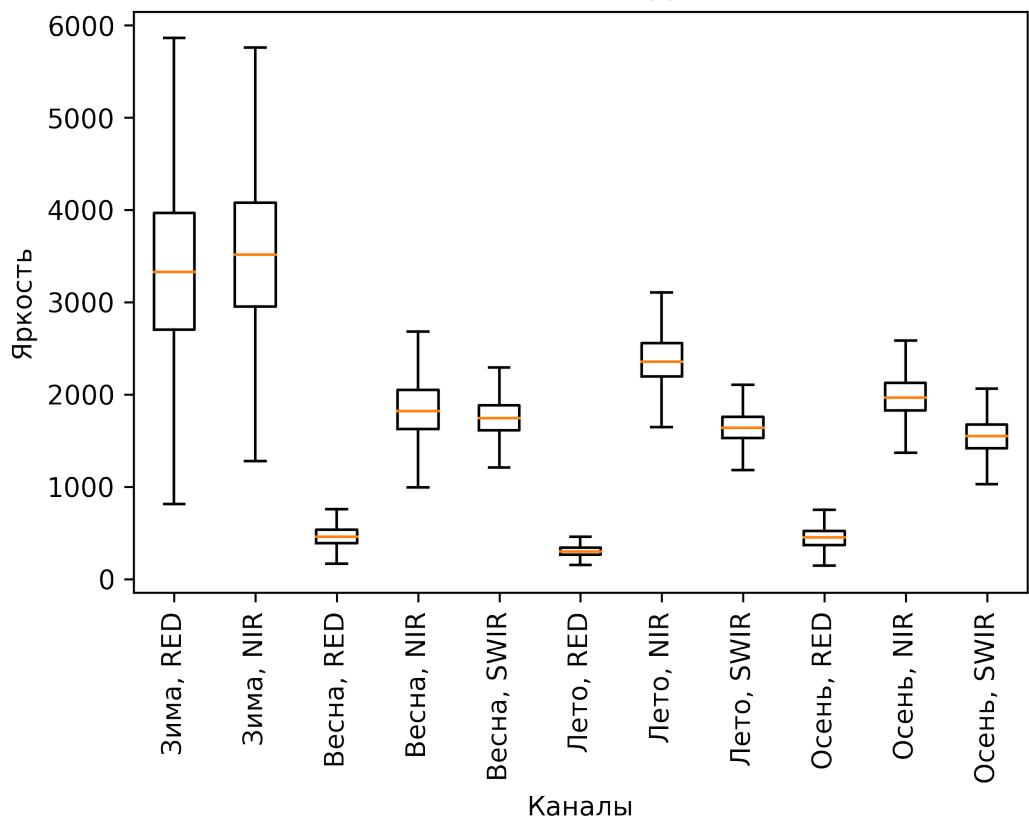
Смешанный лес с преобладанием лиственных



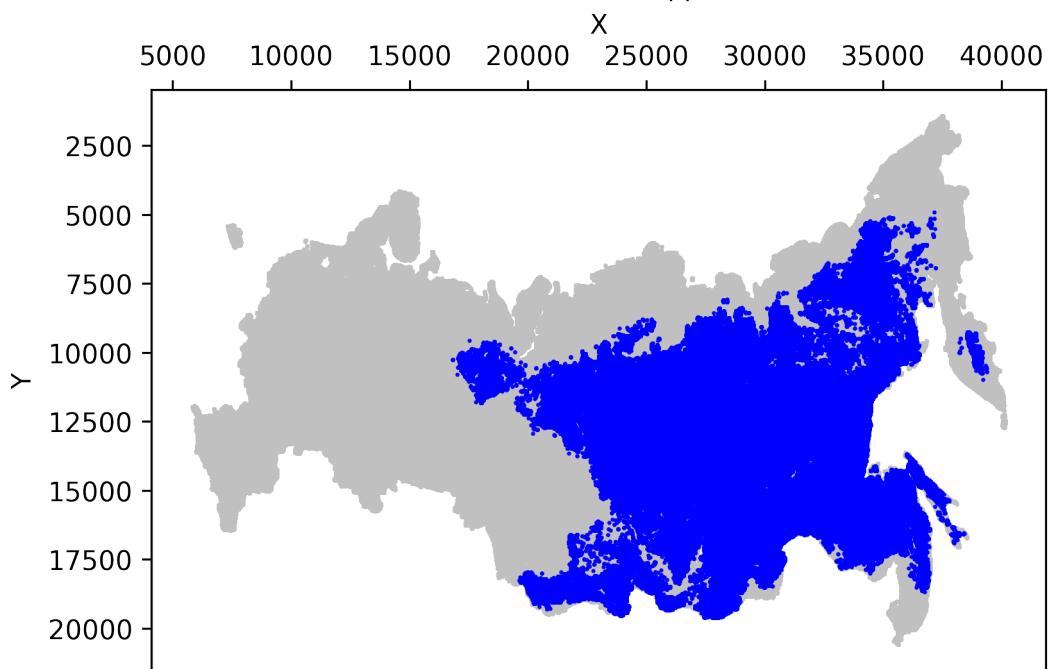
Смешанный лес с преобладанием лиственных



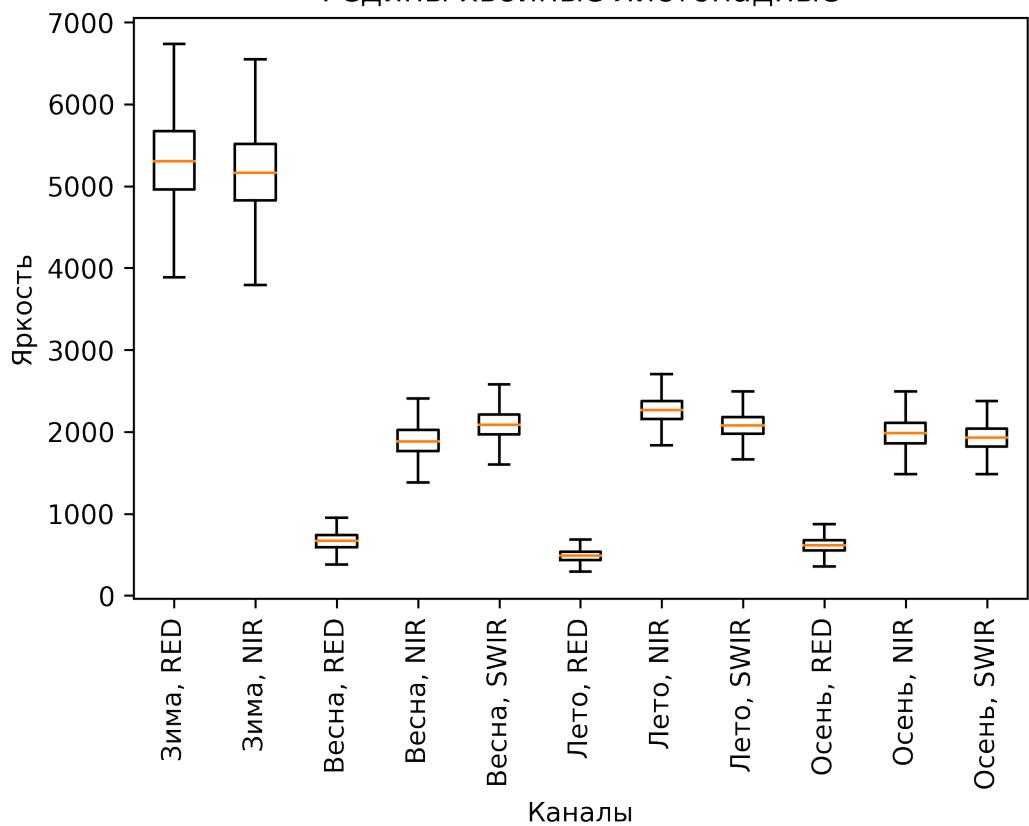
Хвойный листопадный лес



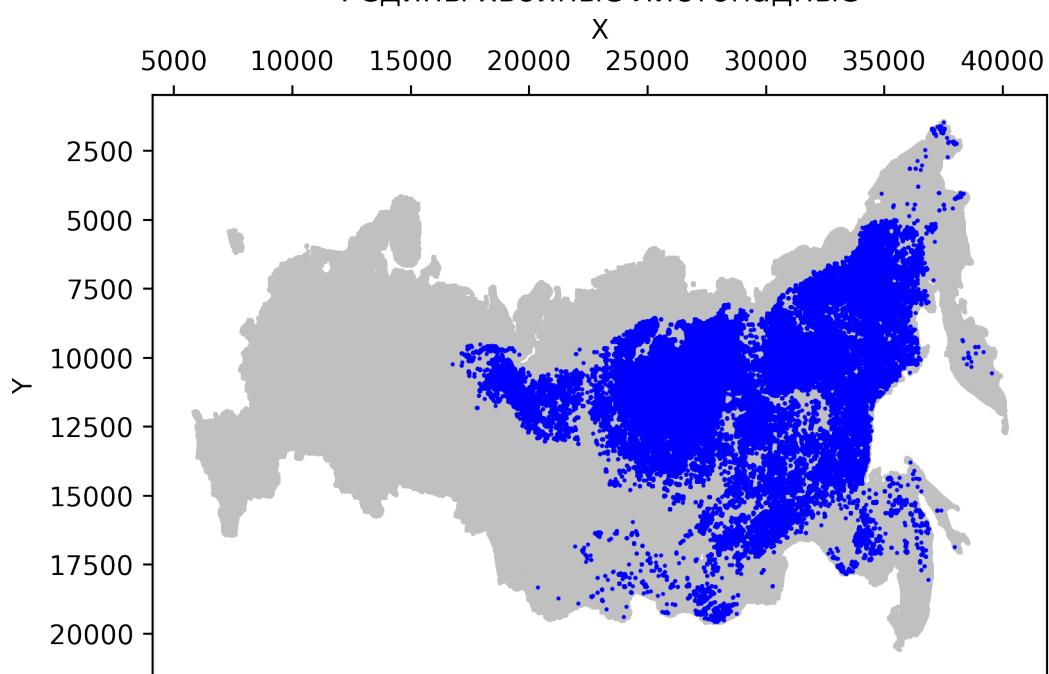
Хвойный листопадный лес



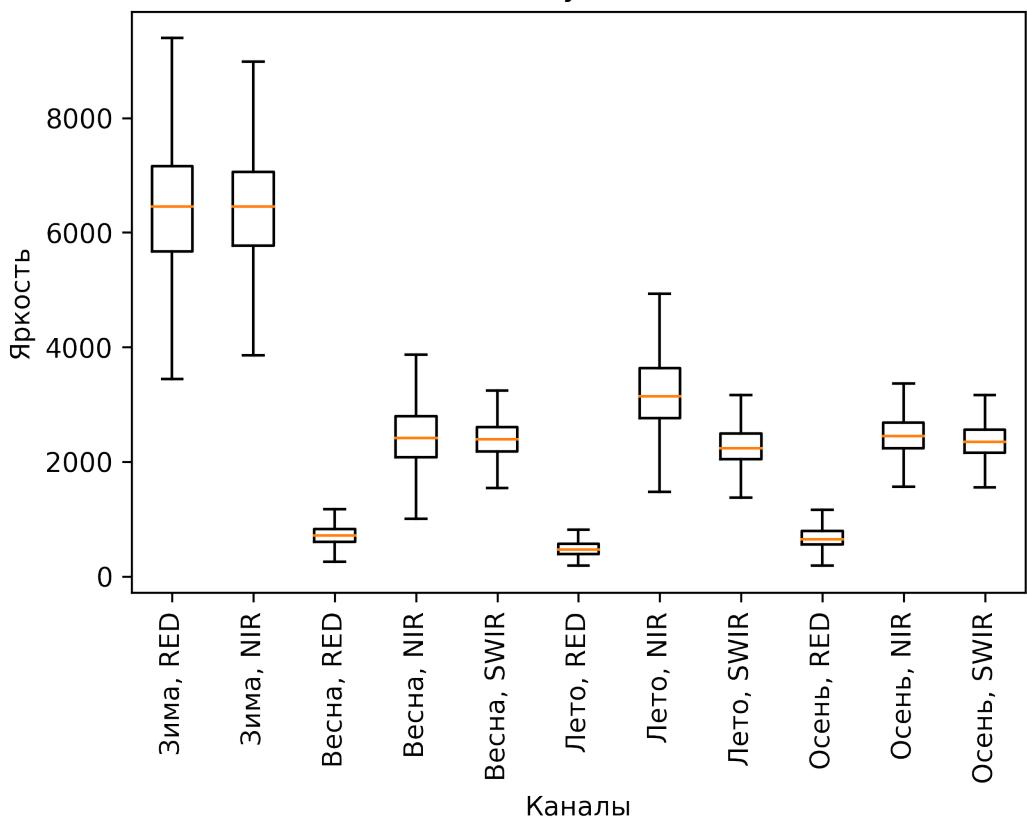
Редины хвойные листопадные



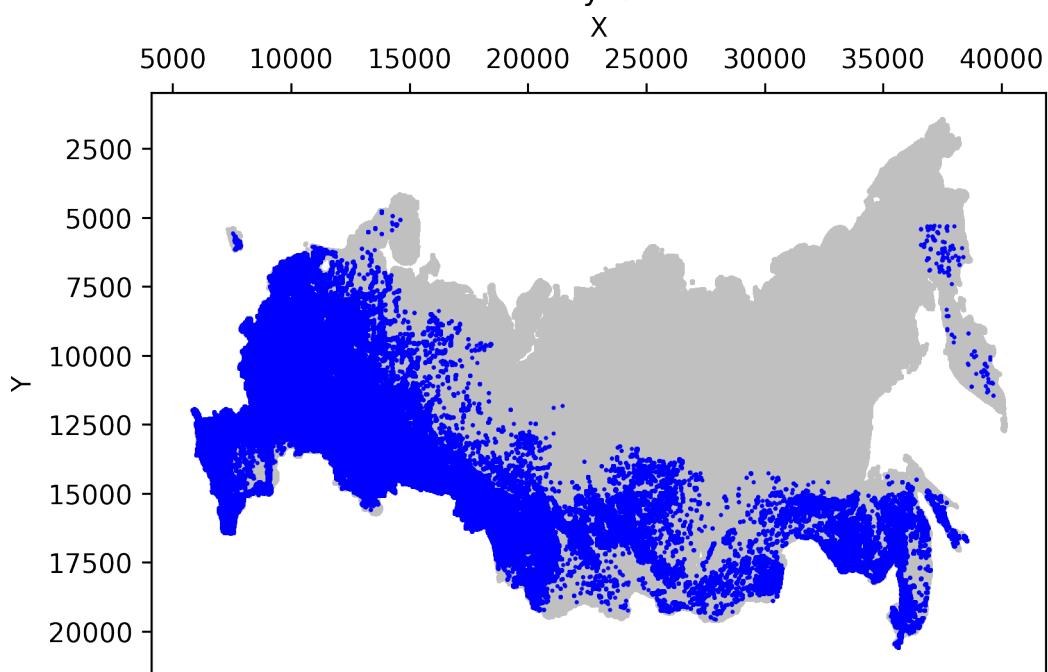
Редины хвойные листопадные

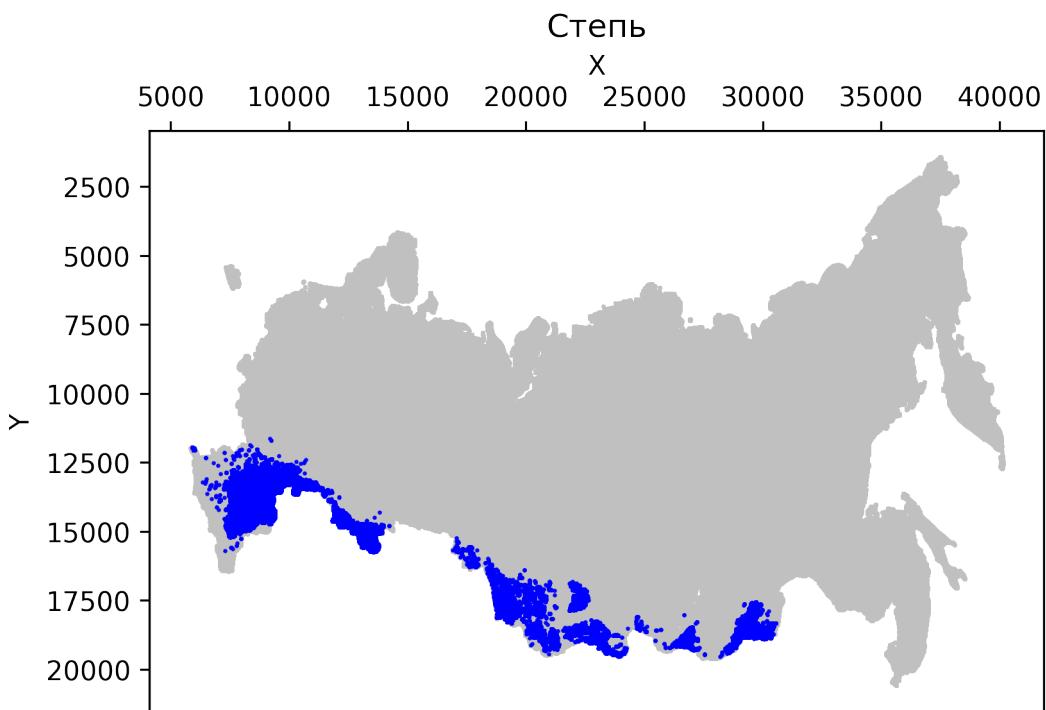
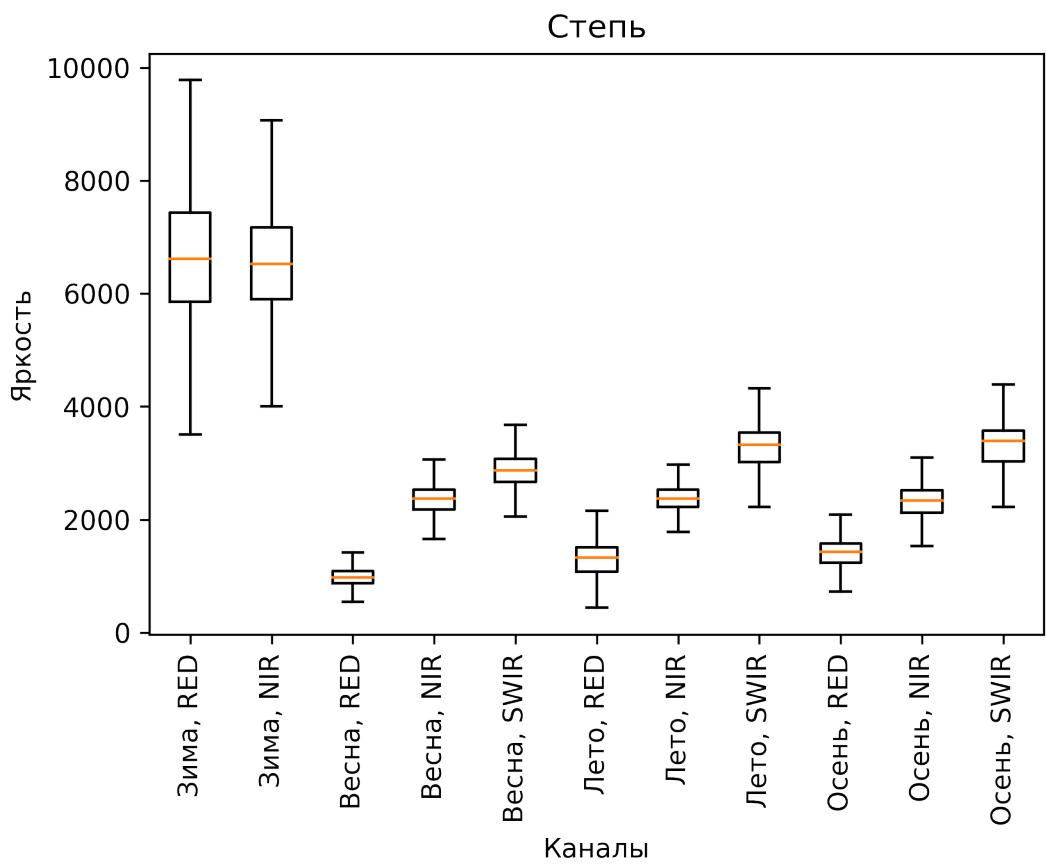


Луга

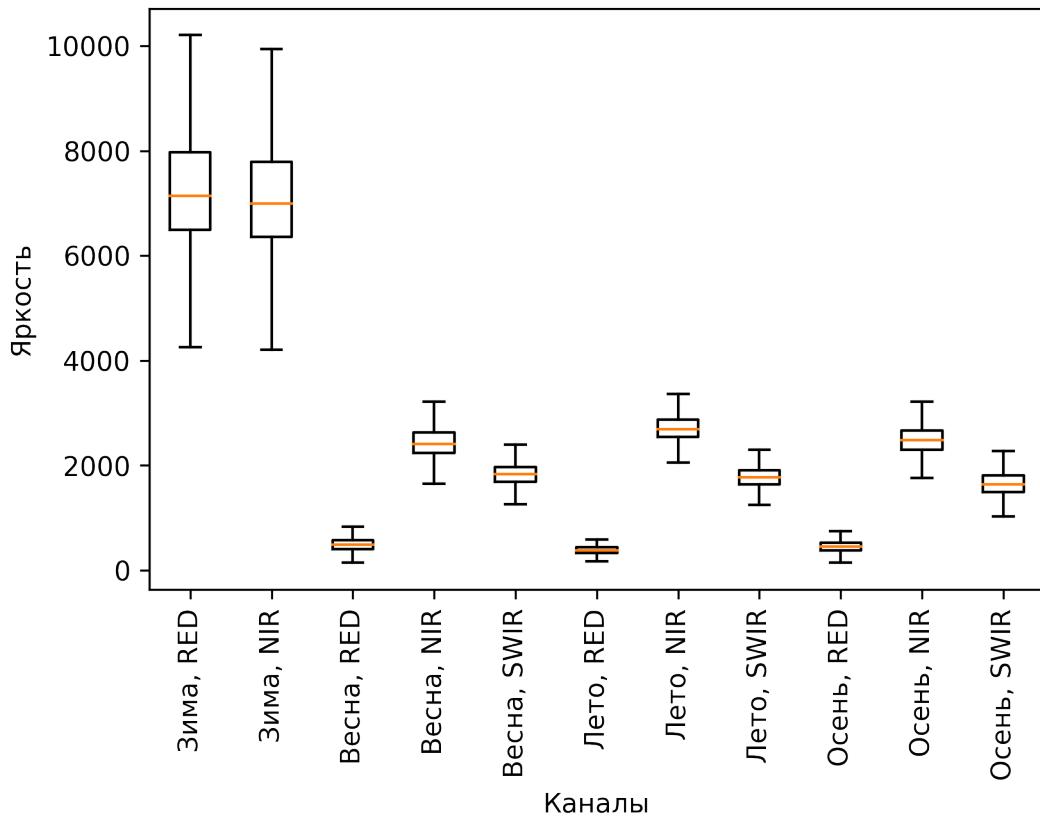


Луга

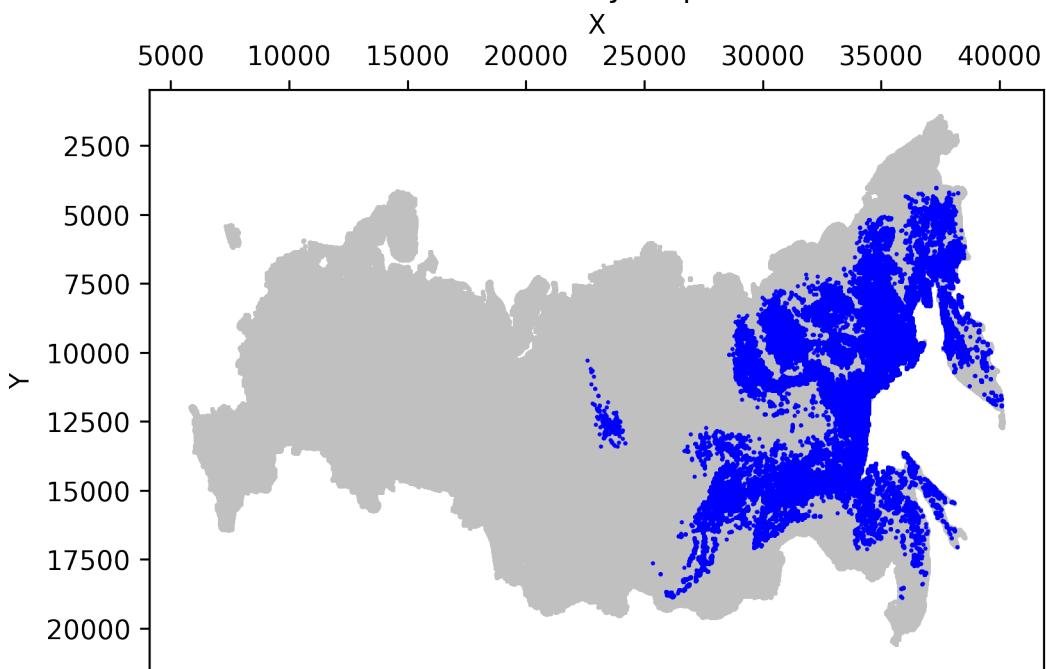




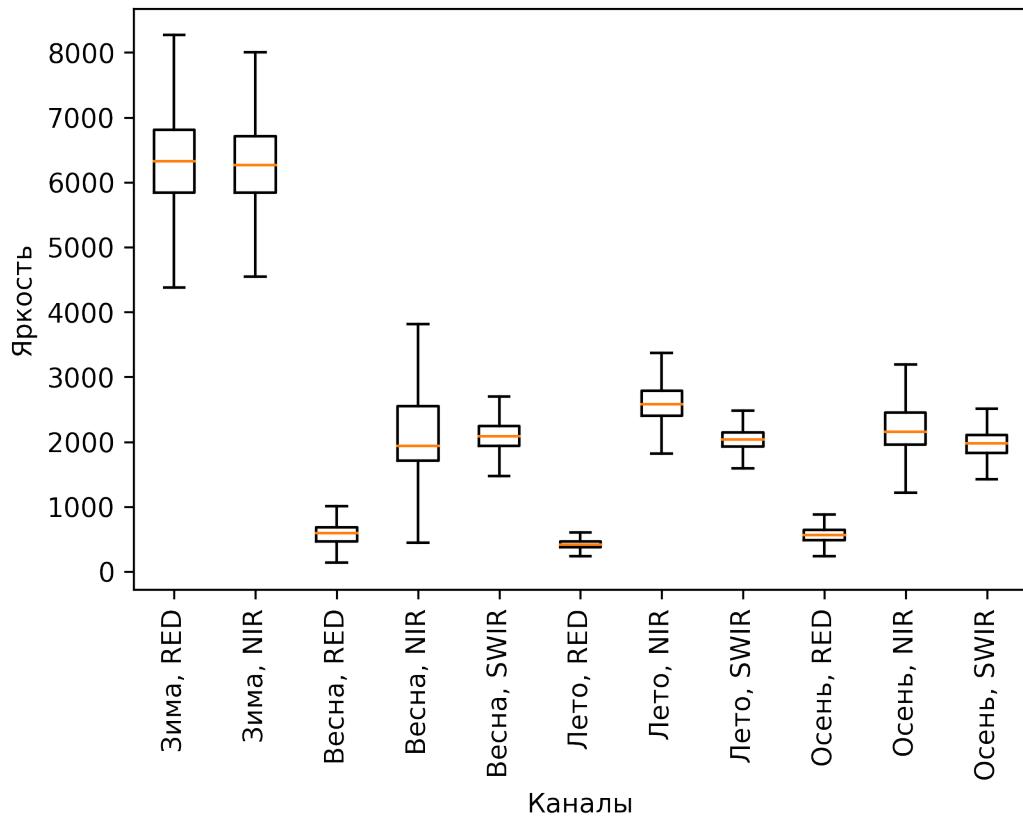
Хвойный кустарник



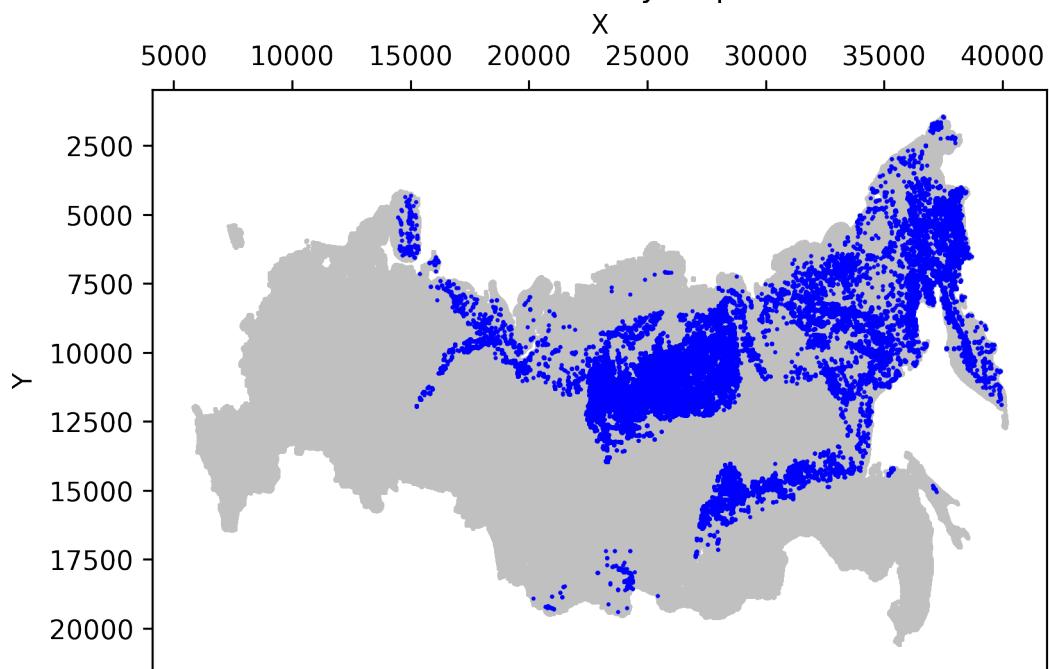
Хвойный кустарник



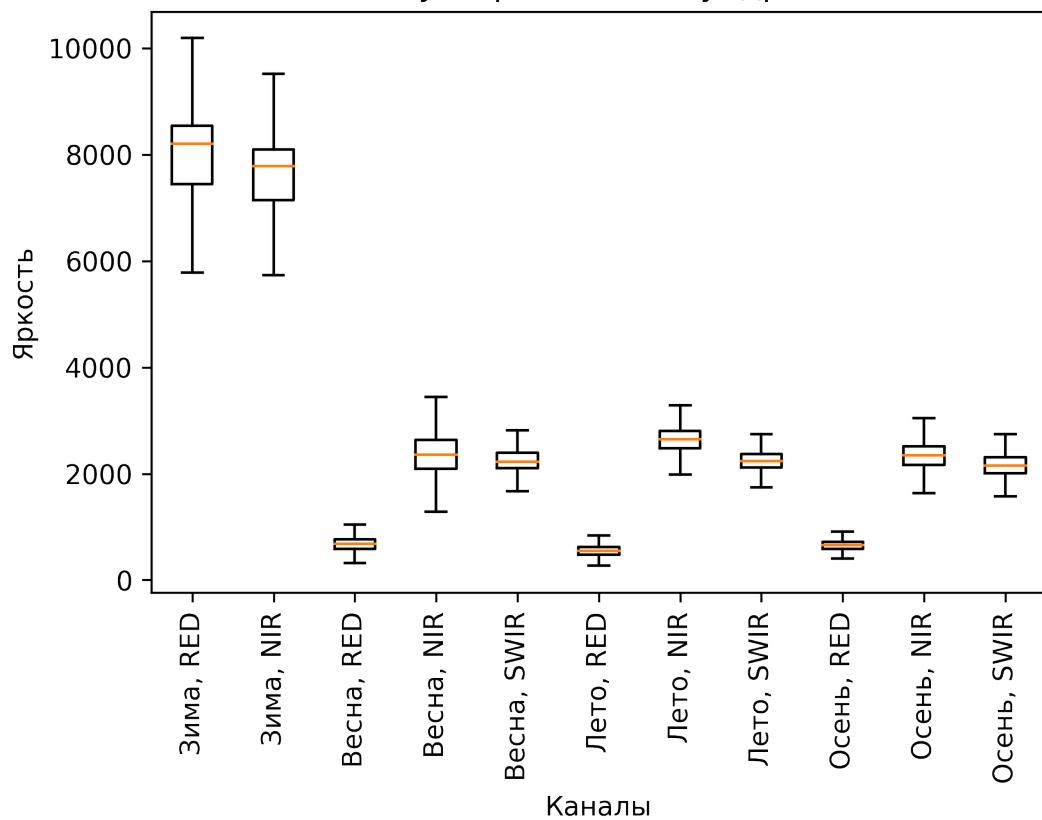
Лиственый кустарник



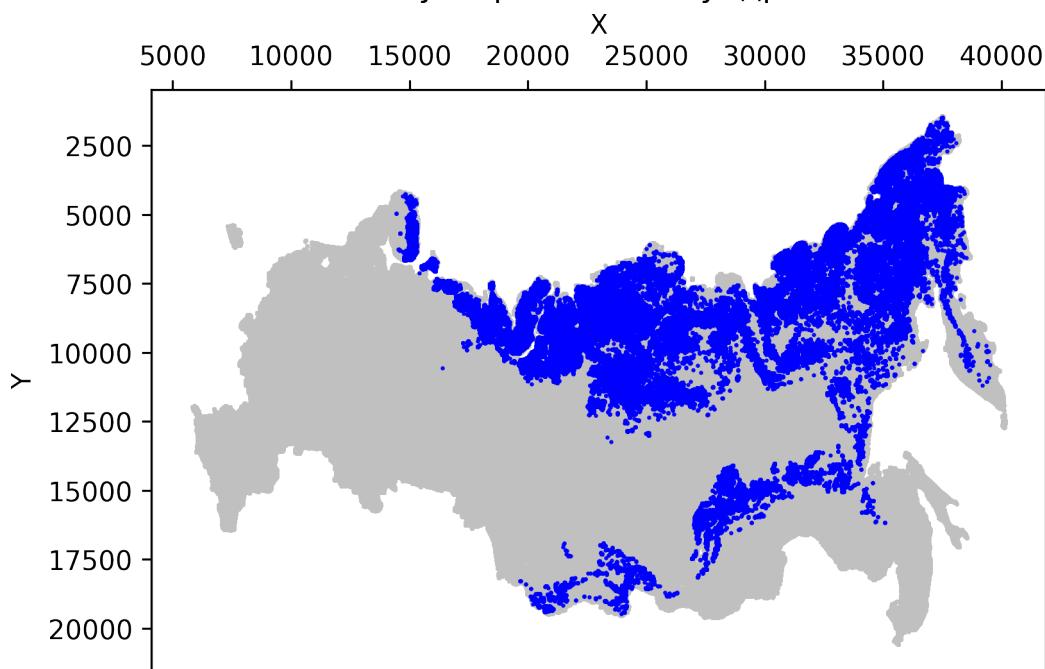
Лиственый кустарник



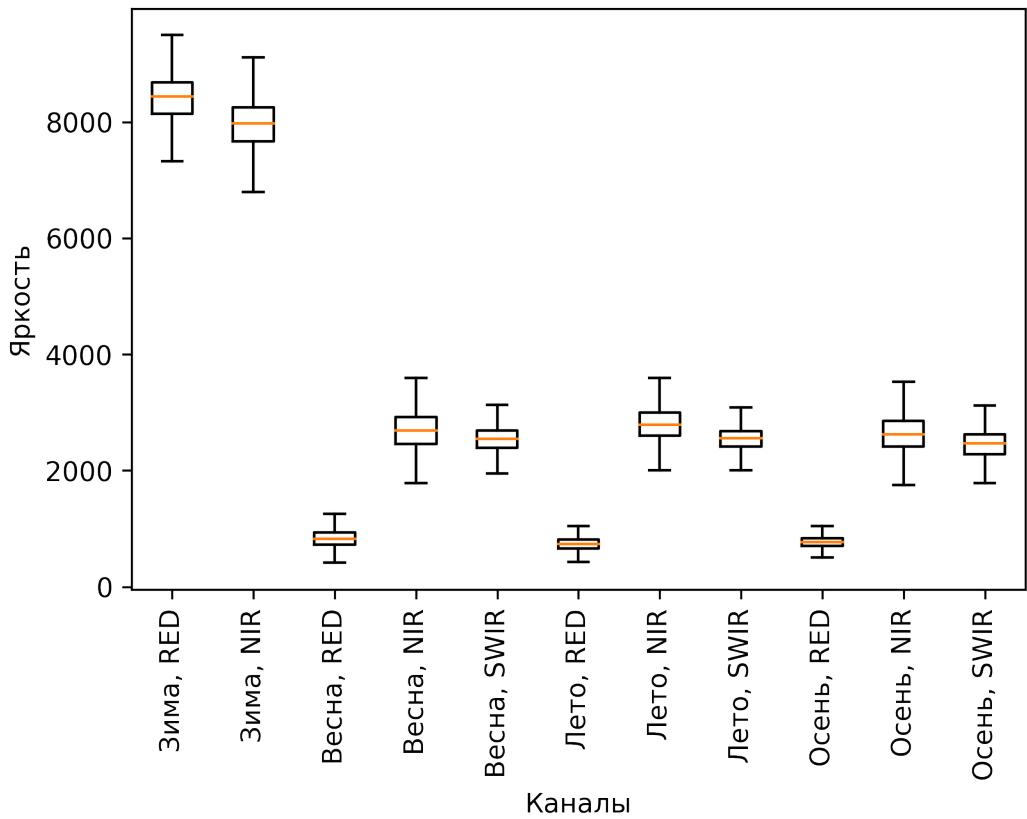
Кустарниковая тундра



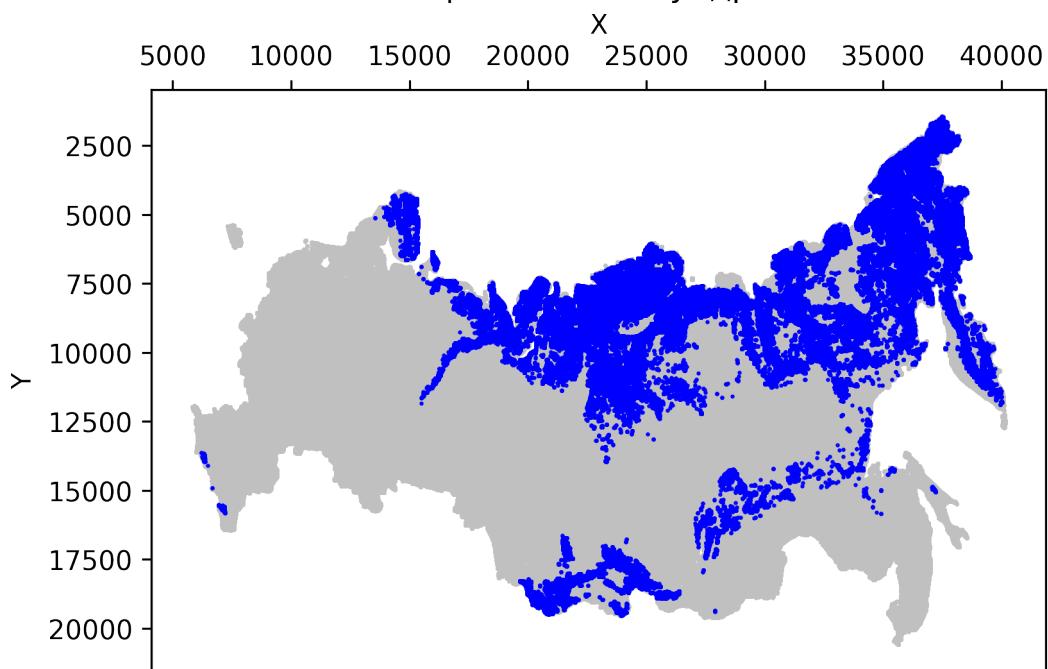
Кустарниковая тундра

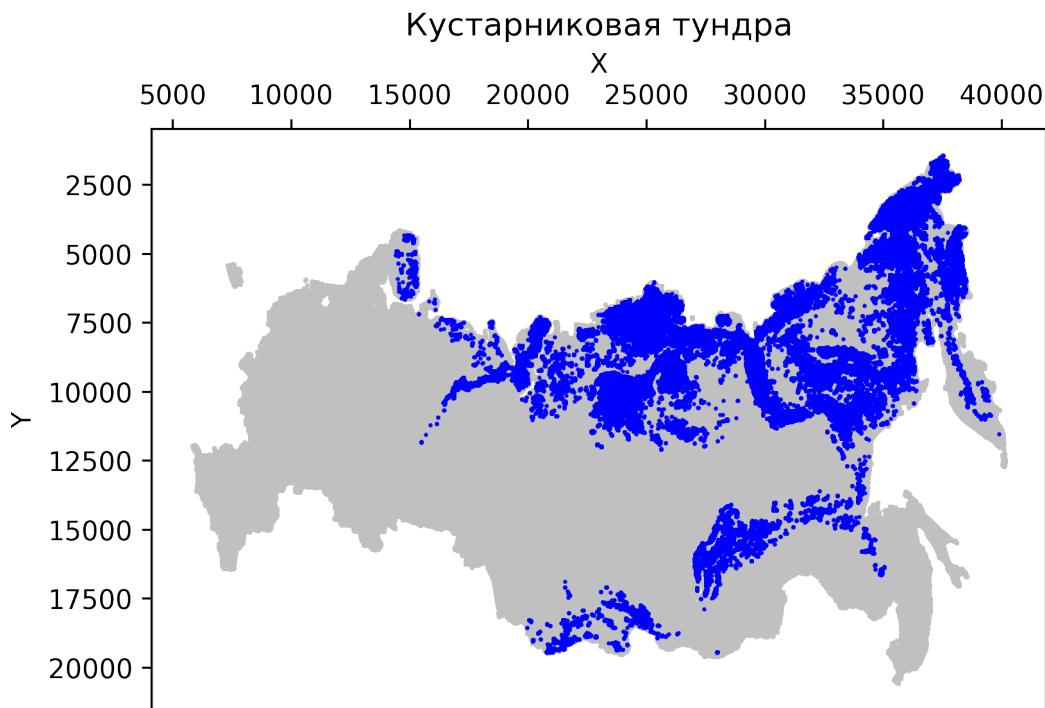
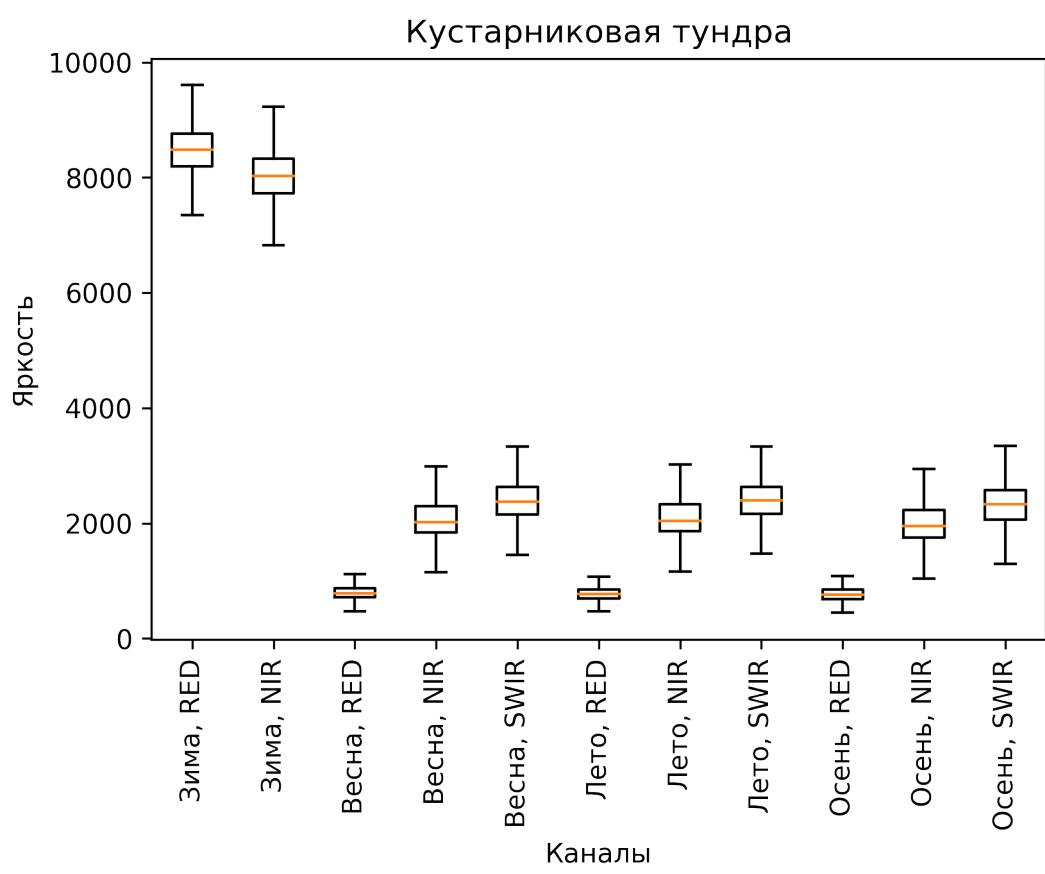


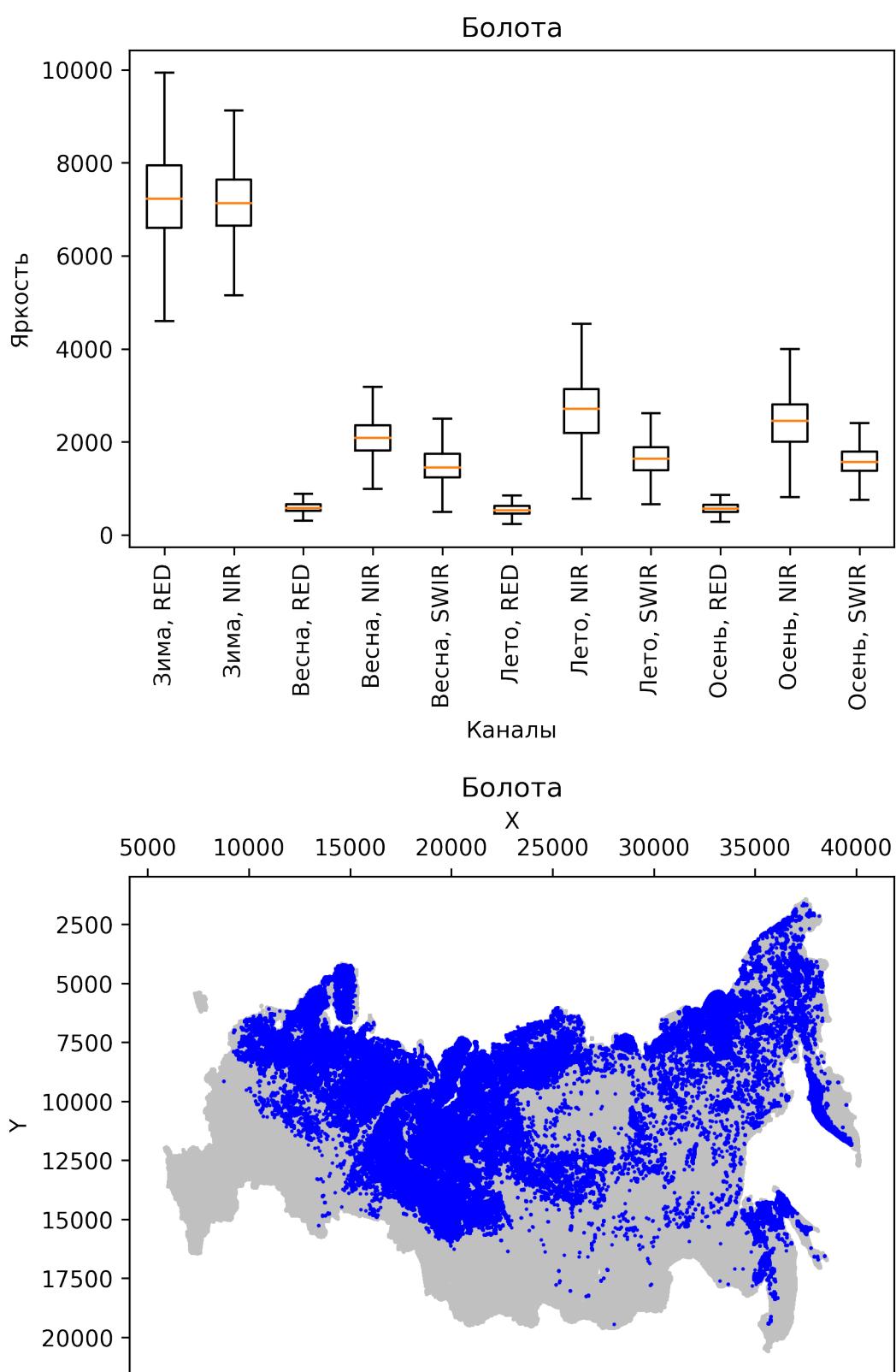
Травянистая тундра



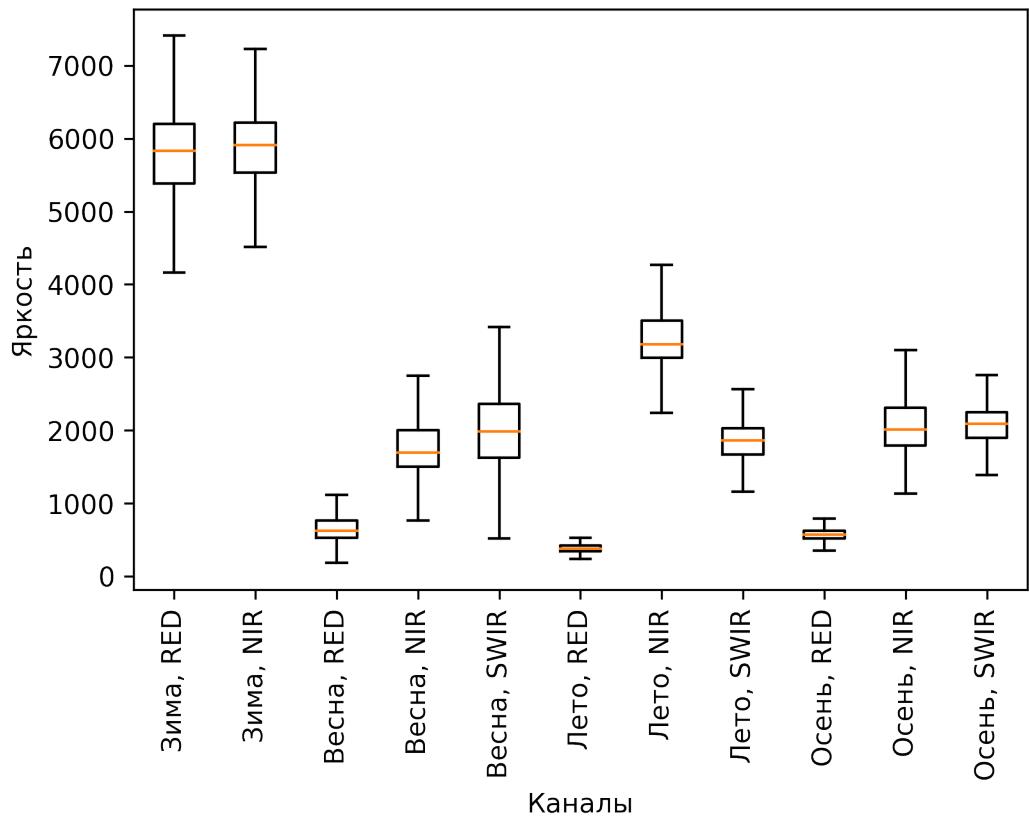
Травянистая тундра



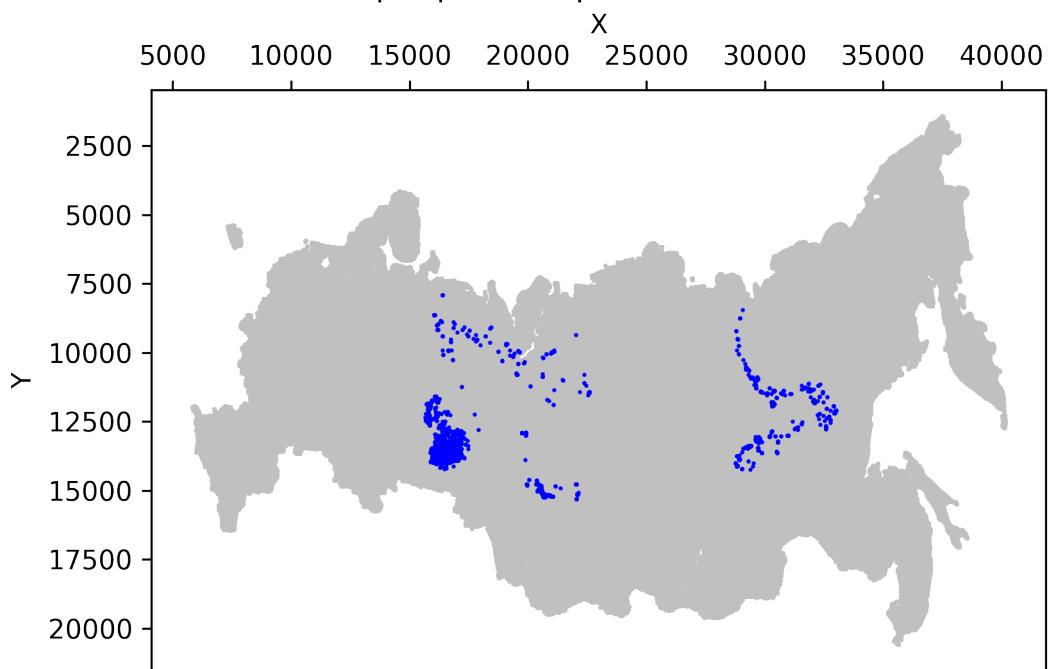




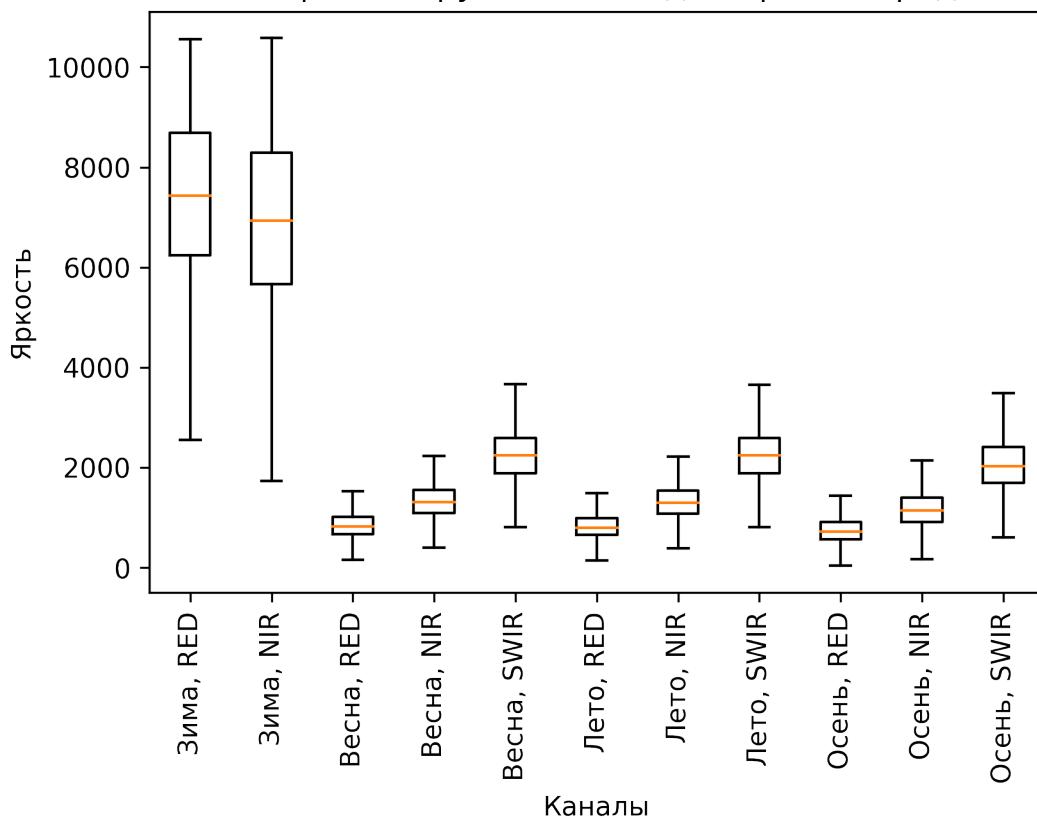
Прибрежная растительность



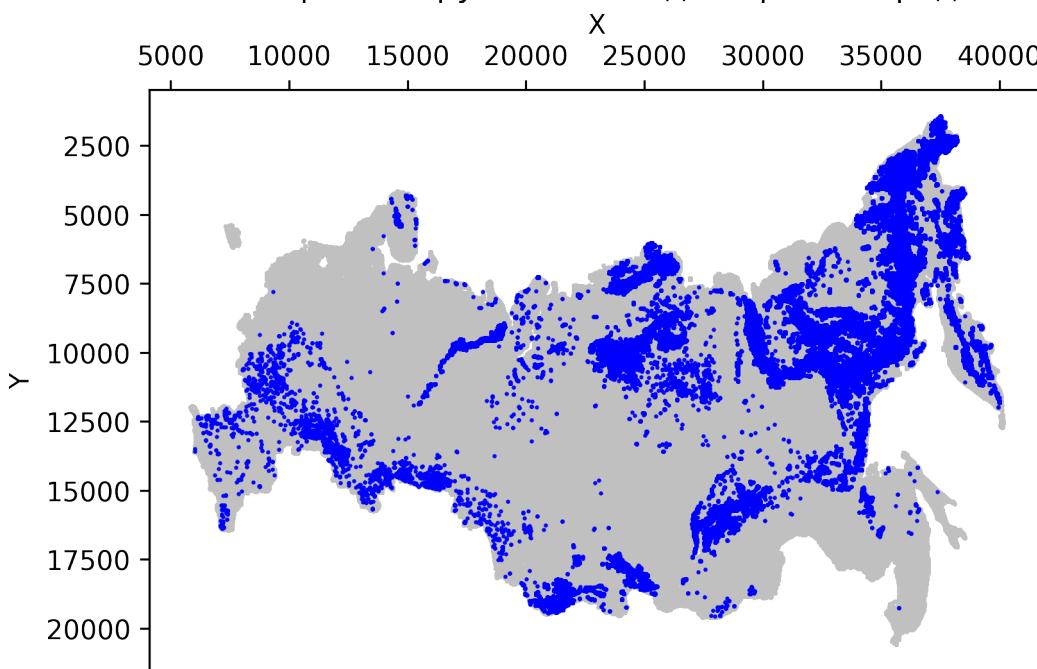
Прибрежная растительность



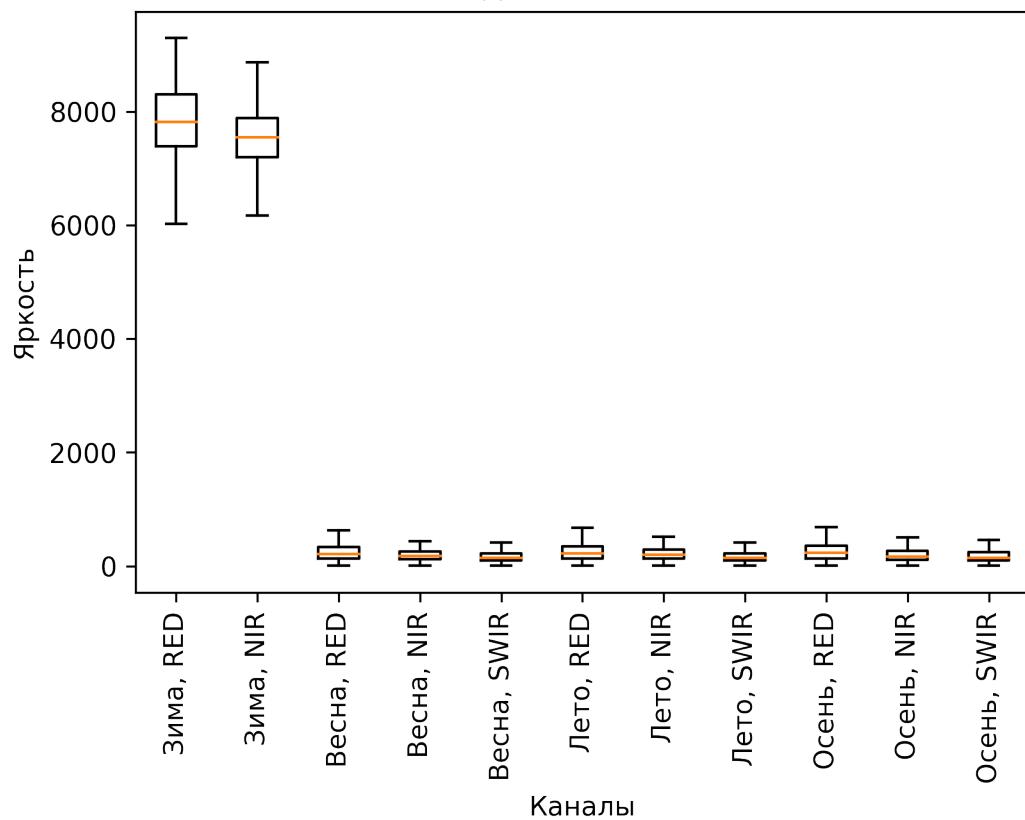
Открытые грунты и выходы горных пород



Открытые грунты и выходы горных пород



Водные объекты



Водные объекты

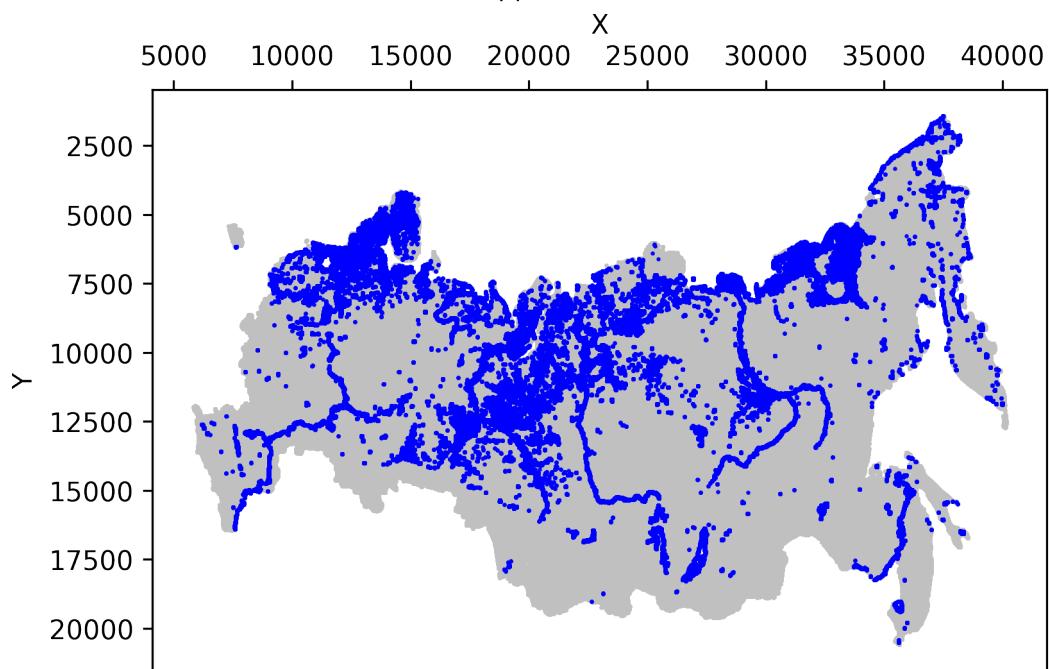


Рис. 3: Сравнение медианных значений сезонных яркостей в различных спектральных каналах по тематическим классам земного покрова, представленным в выборке

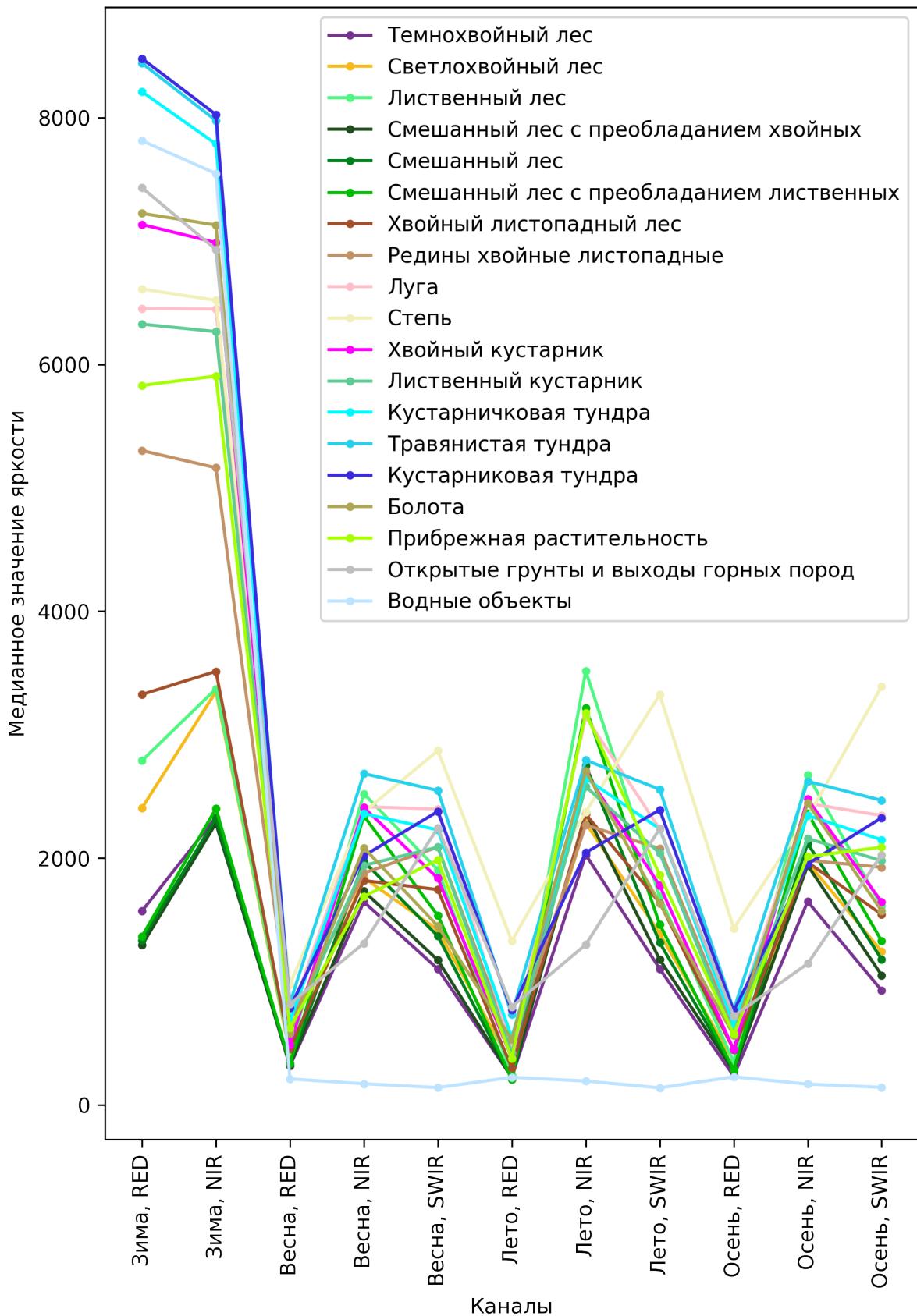
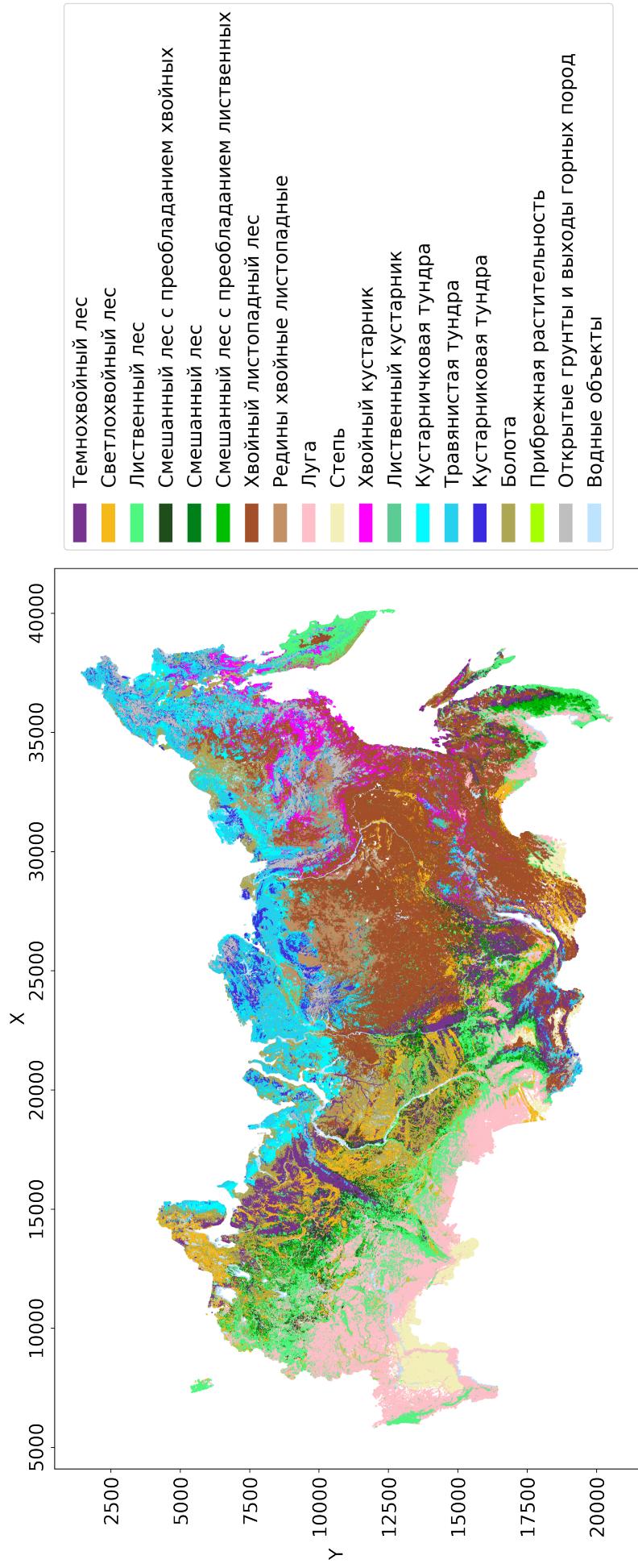


Рис. 4: Карта земного покрова России по данным, предоставленным в выборке



2.4 Инструменты для анализа данных

По той причине, что для анализа данных необходимо работать с полной выборкой, содержащей 74029669 элементов, возникает проблема нехватки оперативной памяти на локальной машине при использовании классического инструмента для анализа данных — библиотеки Pandas [5]. В качестве более производительной альтернативы библиотеке Pandas в настоящей работе для анализа данных применяется библиотека Vaex [6], использующая политику нулевого копирования, параллельные и ленивые вычисления.

Для визуализации данных применяется библиотека Matplotlib [7].

3 Классификация земного покрова с помощью случайного леса

3.1 Математическая постановка задачи классификации

Введем n -мерное пространство параметров (признаков), полученных по данным дистанционного зондирования Земли: $F = \{(f_1, f_2, \dots, f_n)\}$, где $f_i, i = (1, \dots, n)$ — значения отдельных параметров. В нашей задаче признаками являются координаты элементов выборки и значения сезонных яркостей в различных спектральных каналах. Существует подлежащая дистанционной оценке характеристика C , в нашей задаче дискретная — обозначающая типы земного покрова. Также существует неизвестная функция $A : F \rightarrow C$, которая задает отношение между множеством параметров и возможными значениями характеристики. Предположим, что для некоторого подмножества пространства параметров (обучающей выборки) $F_T \in F$ известны соответствующие им значения характеристики, что можно выразить следующим соотношением: $A_T : F_T \rightarrow C$. Также предположим, что существует некоторый метод, позволяющий на основании знаний об известных значениях характеристик и значениях параметров построить приближение к реальной функции $A : F \rightarrow C$. Обозначим этот метод $TRAIN : \{F\} \times \{F_T\} \rightarrow \{\tilde{A}\}$. В нашей задаче функция $TRAIN$ представляет собой процесс построения классификатора по обучающей выборке. Такой способ машинного обучения называется обучением с учителем.

Важной особенностью при этом является способность обучаемой системы к обобщению, то есть к адекватному отклику на данные, выходящие за пределы имеющейся обучающей выборки. Для измерения точности ответов вводится оценочный функционал качества — метрика.

3.2 Численная оценка качества алгоритма

Основой оценки качества работы алгоритма является тестовая выборка, в которой задано соответствие между элементами выборки и их классами. При наличии тестовой выборки достаточно применить классификатор к элементам данной выборки и сравнить его решения с заранее известными экспертными решениями. Однако, для того чтобы делать выводы об относительном качестве работы различных алгоритмов или одного алгоритма с различными параметрами необходима численная метрика качества алгоритма.

Задача классификации земного покрова является задачей мультиклассовой классификации, так как целевая функция (характеристика C) принимает более двух значений.

Многие метрики классификации определены для бинарной классификации и требуют усреднения по классам для получения одной оценки для мультиклассовой классификации. Существует несколько подходов к решению названной проблемы: micro-, macro- и weighted-averaging.

Пусть выборка состоит из K классов. Рассмотрим K задач бинарной классификации, каждая из которых заключается в отделении одного класса от остальных. Для каждой из них можно вычислить различные характеристики (метрики) алгоритма. При подходе micro-averaging метрика вычисляется глобально, путем подсчета итоговых истинно-положительных, истинно-отрицательных, ложно-положительных и ложно-отрицательных результатов (независимо от классов). При подходе macro-averaging сначала вычисляется метрика для каждого класса, затем — невзвешенное среднее по всем классам. При подходе weighted-averaging вычисляется метрика для каждого класса, а затем — взвешенное среднее по числу представителей каждого класса.

В случае неравных по количеству представителей классов важно подбирать метрику, которая будет корректно отражать качество классификации.

Далее приведен обзор метрик, используемых для оценки качества работы алгоритмов в задачах классификации.

3.2.1 Accuracy

Метрика accuracy (точность) определяется долей элементов выборки, по которым классификатор принял правильное решение.

$$Accuracy = \frac{P}{N},$$

где P — количество элементов выборки по которым классификатор принял правильное решение, N — размер обучающей выборки.

Недостатком данной метрики является то, что она присваивает всем элементам выборки одинаковый вес, что может быть не корректно в случае если распределение элементов в обучающей выборке сильно смещено в сторону какого-либо одного или нескольких классов. В этом случае у классификатора есть больше информации по этим классам и соответственно в рамках этих классов он будет принимать более адекватные решения. На практике это приводит к тому, что мы имеем accuracy, допустим, 80%, но при этом в рамках некоторого конкретного класса классификатор работает плохо, не определяя правильно и трети его представителей.

3.2.2 Precision, Recall

Для оценки качества работы алгоритма на каждом из классов по отдельности вводятся метрики precision (точность системы в пределах класса) и recall (полнота).

Точность системы в пределах класса — это доля элементов выборки действительно принадлежащих данному классу относительно всех элементов которые система отнесла к этому классу.

Полнота системы — это доля найденных классификатором элементов выборки принадлежащих классу относительно всех элементов этого класса в тестовой выборке.

Допустим, мы имеем два класса и алгоритм, предсказывающий принадлежность каждого объекта одному из классов, тогда матрица ошибок классификации будет выглядеть следующим образом:

Таблица 2: Матрица ошибок классификации

	$y = 1$	$y = 0$
$\hat{y} = 1$	True Positive (TP)	False Positive (FP)
$\hat{y} = 0$	False Negative (FN)	True Negative (TN)

В таблице содержится информация о том, сколько раз алгоритм принял верное и сколько раз неверное решение для элементов заданного класса (значения y соответствуют экспертной оценке, значения \hat{y} соответствуют оценке алгоритма).

- True Positive (TP) — истинно-положительное решение;
- True Negative (TN) — истинно-отрицательное решение;
- False Positive (FP) — ложно-положительное решение;
- False Negative (FN) — ложно-отрицательное решение.

Тогда, точность и полнота определяются следующим образом:

$$\text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN}.$$

Precision и recall не зависят, в отличие от accuracy, от соотношения классов и потому применимы в условиях несбалансированных выборок.

3.2.3 F-score

На практике максимальная точность и полнота не достижимы одновременно, поэтому, хотелось бы иметь некую метрику, которая объединяла бы в себе информацию о точности и полноте алгоритма.

F-score представляет собой гармоническое среднее между точностью и полнотой. Она стремится к нулю, если точность или полнота стремятся к нулю.

$$F\text{-score} = 2 \frac{Precision \cdot Recall}{Precision + Recall}$$

Данная формула придает одинаковый вес точности и полноте, поэтому F-score будет уменьшаться одинаково при уменьшении и точности, и полноты.

Возможно рассчитать F-score, придав различный вес точности и полноте, если отдается приоритет одной из этих метрик при разработке алгоритма.

$$F\text{-score} = (\beta^2 + 1) \frac{Precision \cdot Recall}{\beta^2 Precision + Recall},$$

где $0 < \beta < 1$, если приоритет отдается точности, $\beta > 1$, если приоритет отдается полноте. При $\beta = 1$ формула сводится к предыдущей и мы получаем сбалансированную F-score (также ее называют F_1).

F-score является хорошим кандидатом на формальную метрику оценки качества классификатора.

3.2.4 Confusion Matrix

Confusion matrix (матрица ошибок, матрица неточностей) позволяет наглядно представить результаты работы классификатора в случае если количество классов относительно невелико (не более 100-150).

Матрица ошибок — это матрица размера $N \times N$, где N — количество классов. Столбцы матрицы резервируются за экспертными решениями, а строки — за решениями классификатора. При классификации элемента из тестовой выборки мы инкрементируем число, стоящее на пересечении строки класса который вернул классификатор и столбца класса, к которому действительно относится элемент.

При наличии матрицы ошибок точность алгоритма в пределах класса рассчитывается как отношение соответствующего этому классу диагонального элемента матрицы и суммы значений всей строки класса,

полнота — как отношение диагонального элемента матрицы и суммы значений всего столбца класса.

$$Precision_c = \frac{a_{c,c}}{\sum_{i=1}^n a_{c,i}},$$

$$Recall_c = \frac{a_{c,c}}{\sum_{i=1}^n a_{i,c}},$$

где $a_{i,j}$ — значение i -го по строке, j -го по столбцу элемента матрицы, c — индекс класса, n — количество классов.

Результирующая точность классификатора рассчитывается как арифметическое среднее его точности по всем классам, аналогично рассчитывается полнота, данный подход называется macro-averaging.

3.2.5 Receiver Operating Characteristic Curve, Area Under Receiver Operating Characteristic Curve

Receiver operating characteristic curve (ROC, кривая ошибок) является одним из способов оценить модель в целом и представляет из себя линию от точки $(0, 0)$ до точки $(1, 1)$ в координатах True Positive Rate (TPR) и False Positive Rate (FPR):

$$TPR = \frac{TP}{TP + FN},$$

$$FPR = \frac{FP}{FP + TN},$$

где TPR — полнота (доля элементов класса 1, которые верно klassифицированы алгоритмом), FPR — доля элементов класса 0, которые неверно klassифицированы алгоритмом.

Area Under Receiver Operating Characteristic Curve (AUC ROC, площадь под кривой ошибок) является агрегированной характеристикой качества klassификации, не зависящей от соотношения цен ошибок. Чем больше значение AUC ROC, тем лучше модель klassификации, кроме этого, важной является крутизна кривой — мы максимизируем TPR, минимизируя FPR, а значит, кривая должна стремиться к точке $(0, 1)$.

В идеальном случае, когда классификатор не делает ошибок ($FPR = 0$, $TPR = 1$) мы получим площадь под кривой, равную 1; в случае, когда классификатор случайно выдает метки классов, AUC ROC будет стремиться к 0,5, так как классификатор будет выдавать одинаковое количество TP и FP.

Критерий AUC ROC устойчив к несбалансированным классам.

3.2.6 Out-of-Bag Error

Out-of-bag error (OOB) — это метод измерения ошибки предсказания случайных лесов и других моделей машинного обучения, использующих bagging.

Bagging — технология классификации, использующая композиции алгоритмов (для модели случайного леса это ансамбль решающих деревьев), каждый из которых обучается независимо. Классификаторы не исправляют ошибки друг друга, а компенсируют их при голосовании. Bagging позволяет снизить процент ошибки классификации в случае, когда высока дисперсия ошибки базового метода.

Суть метода измерения ошибки OOB заключается в том, что при использовании технологии bagging каждый классификатор ансамбля обучается не на полной выборке, а на ее части, оставшаяся же часть выборки используется для оценки данного классификатора. Таким образом производится оценка каждого решающего дерева случайного леса, что позволяет получить характеристику качества работы случайного леса в целом.

Ошибка OOB может быть использована для валидации, что позволит подобрать оптимальное количество деревьев случайного леса.

3.3 Случайный лес

В настоящей работе классификация типов земного покрова по данным дистанционного зондирования Земли производится с помощью мо-

дели случайного леса. Используется реализация случайного леса, предоставляемая библиотекой Scikit-learn [8].

3.3.1 Модель случайного леса

Случайный лес — алгоритм машинного обучения, заключающийся в использовании ансамбля решающих деревьев.

Пусть обучающая выборка состоит из N элементов, размерность пространства признаков равна M , и задан параметр t (в задачах классификации обычно $t \approx \sqrt{M}$) как неполное количество признаков для обучения.

Построение деревьев ансамбля (bagging) производится следующим образом:

1. Сгенерируем случайную повторную подвыборку размером N из обучающей выборки. Некоторые элементы попадут в неё два или более раза, тогда как в среднем $N(1 - \frac{1}{N})^N$ (при больших N примерно $\frac{N}{e}$, где e — основание натурального логарифма) образцов оказываются не вошедшими в набор или неотобранными (out-of-bag).
2. Построим решающее дерево, классифицирующее элементы данной подвыборки, причём в ходе создания очередного узла дерева будем выбирать набор признаков, на основе которых производится разбиение (не из всех M признаков, а лишь из t случайно выбранных).
3. Дерево строится до полного исчерпания подвыборки и не подвергается процедуре прунинга (отсечения ветвей).

Классификация элементов проводится путём голосования: каждое дерево ансамбля относит классифицируемый элемент к одному из классов, а побеждает класс, за который проголосовало наибольшее число деревьев.

Увеличение количества деревьев в ансамбле позволяет повысить качество классификации. Однако, при увеличении числа деревьев возрастает так же время обучения и работы случайного леса.

3.3.2 Предпосылки использования случайного леса

Модель случайного леса успешно справляется с классификацией данных высокой размерности, оставаясь при этом быстрой и нечувствительной к переобучению. Оценка значимости признаков, обеспечивающая классификатором, может быть использована для уменьшения числа признаков в выборке, для определения наиболее релевантных данных дистанционного зондирования, а также для выбора наиболее подходящих сезонов и спектральных каналов для классификации определенных типов земного покрова, что показано в работе [3].

3.3.3 Обучение случайного леса и оценка качества классификации

Для обучения случайного леса набор данных разделен на обучающую (2000000 элементов) и тестовую выборки (72029669 элементов). Затем обучающая выборка разделена на непосредственно обучающую (200000 элементов) и валидационную (1800000 элементов) выборки.

Предварительная обработка данных перед обучением модели не осуществляется, так как она не повышает качества работы случайного леса, что было установлено эмпирически, и данные не содержат пропущенных значений.

Произведено обучение случайного леса с гиперпараметрами по умолчанию и получены предсказания типов земного покрова для валидационной выборки. Ниже приведены оценки качества классификации валидационной выборки обученной моделью случайного леса с гиперпараметрами по умолчанию:

$$Accuracy = 0.95334,$$

$F_1\text{-score} = 0.95249$ (при взвешенном усреднении),

$Out-of\text{-bag error} = 0.95189$.

Рис. 5: Матрица ошибок классификации случайного леса с гиперпараметрами по умолчанию

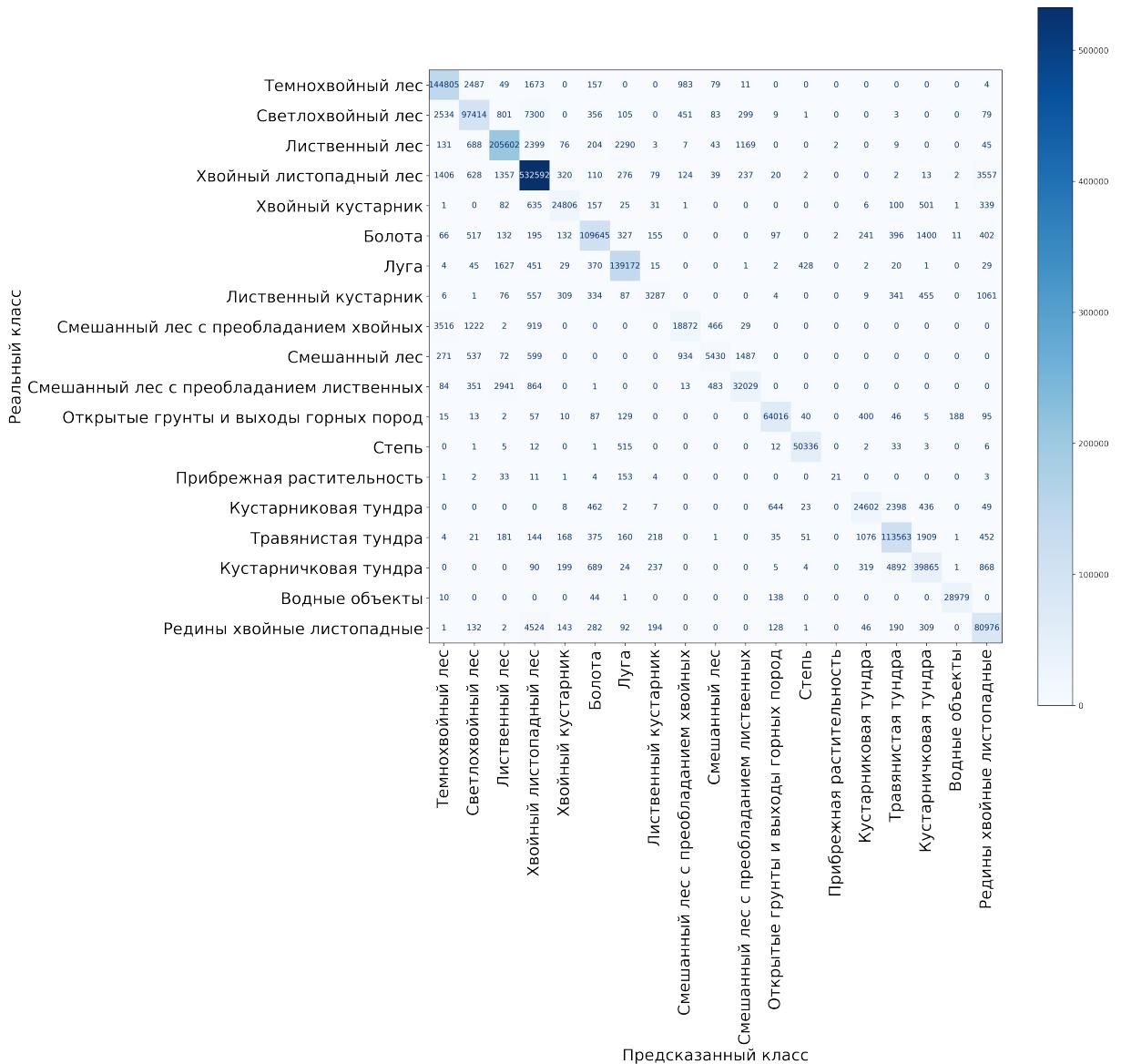
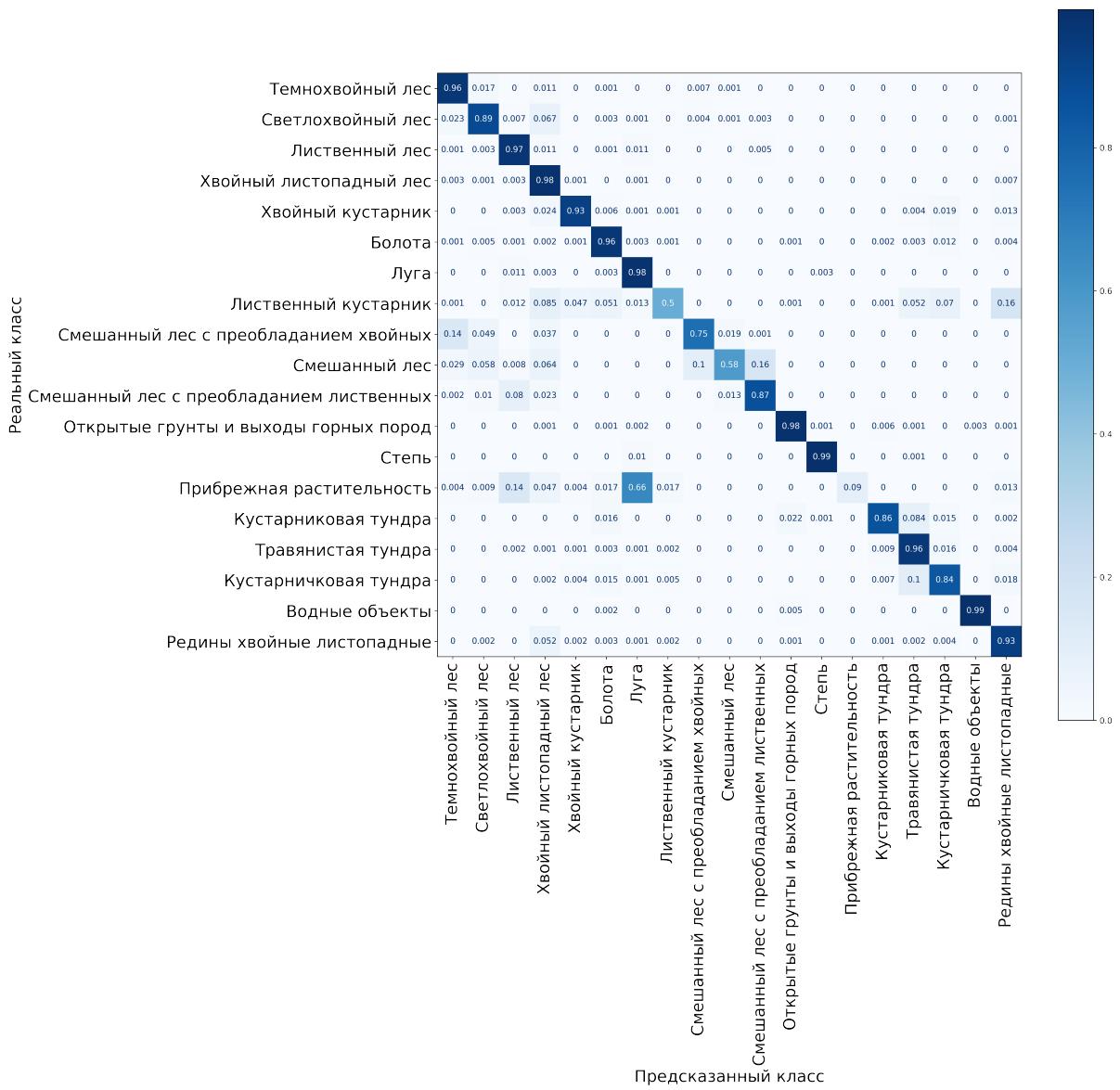


Рис. 6: Нормированная по строкам матрица неточностей классификации случайного леса с гиперпараметрами по умолчанию



3.3.4 Оптимизация гиперпараметров алгоритма

Осуществлён подбор оптимальных гиперпараметров (количество деревьев, числа признаков для выбора расщепления, максимальной глубины дерева, критерия расщепления) модели случайного леса для данной выборки.

Оптимальное количество деревьев — 600, число признаков для выбора расщепления — квадратный корень количества признаков, максимальная глубина дерева — 20, критерий расщепления — энтропийный.

Однако, качество классификации моделью с оптимальными гиперпараметрами возрастает крайне несущественно ($F_1\text{-score} = 0.95342$ (при взвешенном усреднении), что лишь на тысячные доли превышает значение $F_1\text{-score} = 0.95249$ модели с гиперпараметрами по умолчанию). При этом время обучения модели ощутимо возрастает (в основном из-за увеличения количества и глубины деревьев). Поэтому принято решение использовать для классификации модель с частично оптимальными гиперпараметрами (число признаков для выбора расщепления — квадратный корень количества признаков, критерий расщепления — энтропийный), для которой $F_1\text{-score} = 0.95341$.

3.3.5 Проверка случайного леса на полной тестовой выборке

Произведено обучение случайного леса с выбранными гиперпараметрами на полной обучающей выборке (2000000 элементов) и получены предсказания типов земного покрова для полной тестовой выборки (72029669 элементов). Ниже приведены оценки качества классификации полной тестовой выборки обученной моделью случайного леса с выбранными гиперпараметрами:

$$Accuracy = 0.97214,$$

$$F_1\text{-score} = 0.97185 \text{ (при взвешенном усреднении)},$$

$$Out-of-bag error = 0.97118.$$

Рис. 7: Матрица ошибок классификации случайного леса с выбранными гиперпараметрами

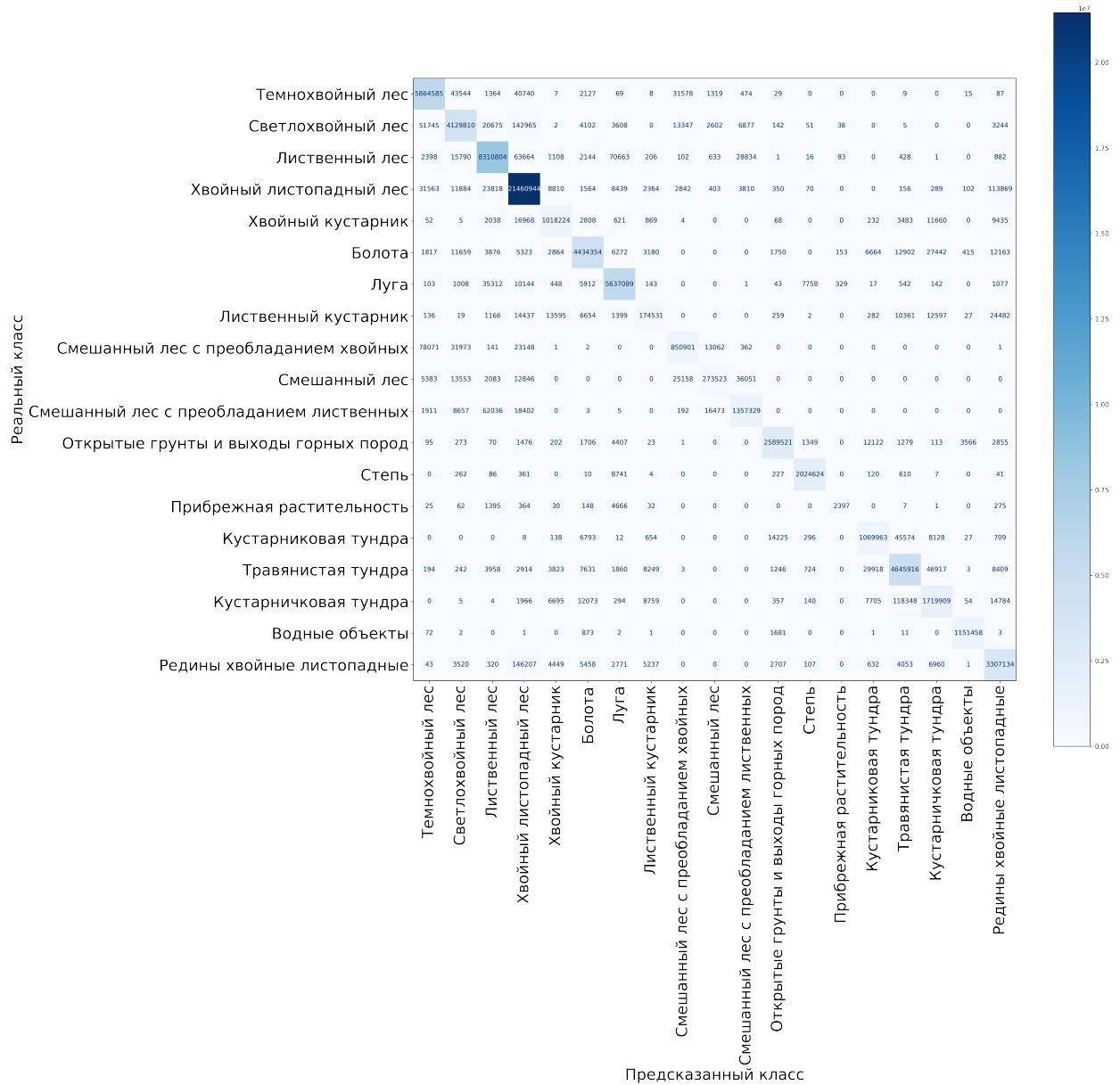
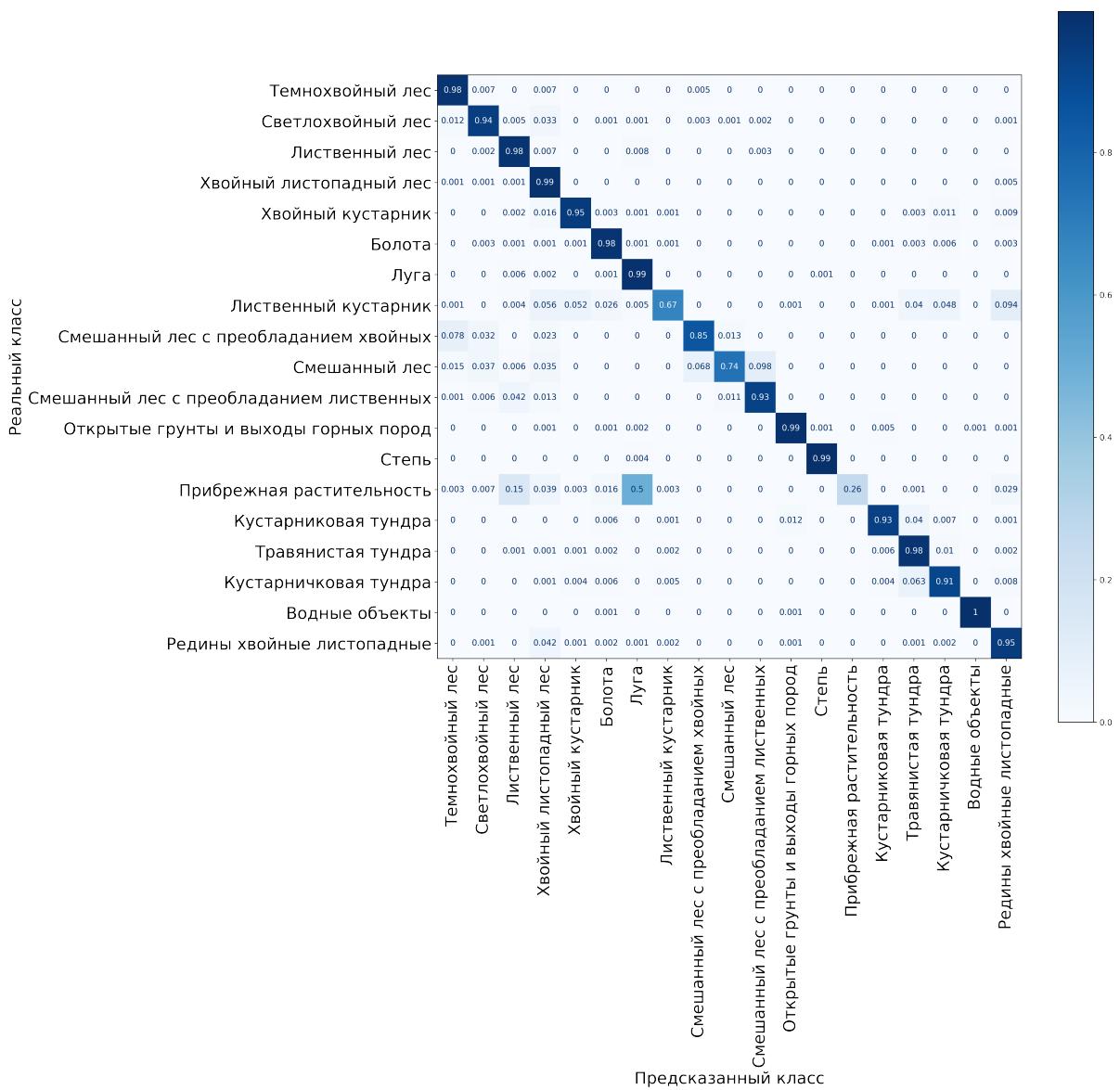


Рис. 8: Нормированная по строкам матрица неточностей
классификации случайного леса с выбранными гиперпараметрами

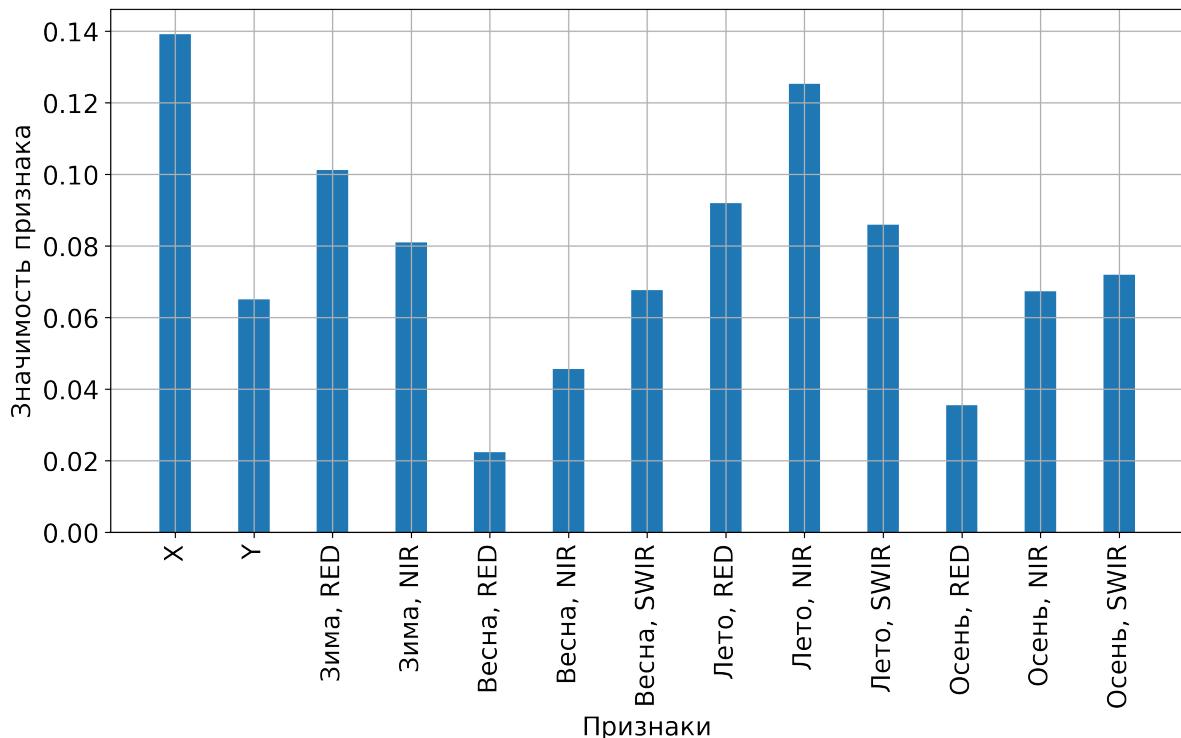


4 Оценка значимости признаков

4.1 Оценка значимости признаков на основе информации, предоставляемой моделью случайного леса

Алгоритм случайного леса использует ансамбль деревьев, которые содержат узлы, полученные в результате расщепления. Основная цель расщеплений — максимальное уменьшение количества шумов таких, как энтропия и коэффициент Джини. Деревья предоставляют информацию о важности признаков, рассчитывая степень минимизации шумов за счёт каждого признака.

Рис. 9: Оценка значимости признаков на основе информации, предоставляемой моделью случайного леса



Согласно информации, предоставляемой моделью случайного леса, отсортированная по убыванию значимости последовательность признаков выглядит следующим образом: X, SUMMER2, WINTER1,

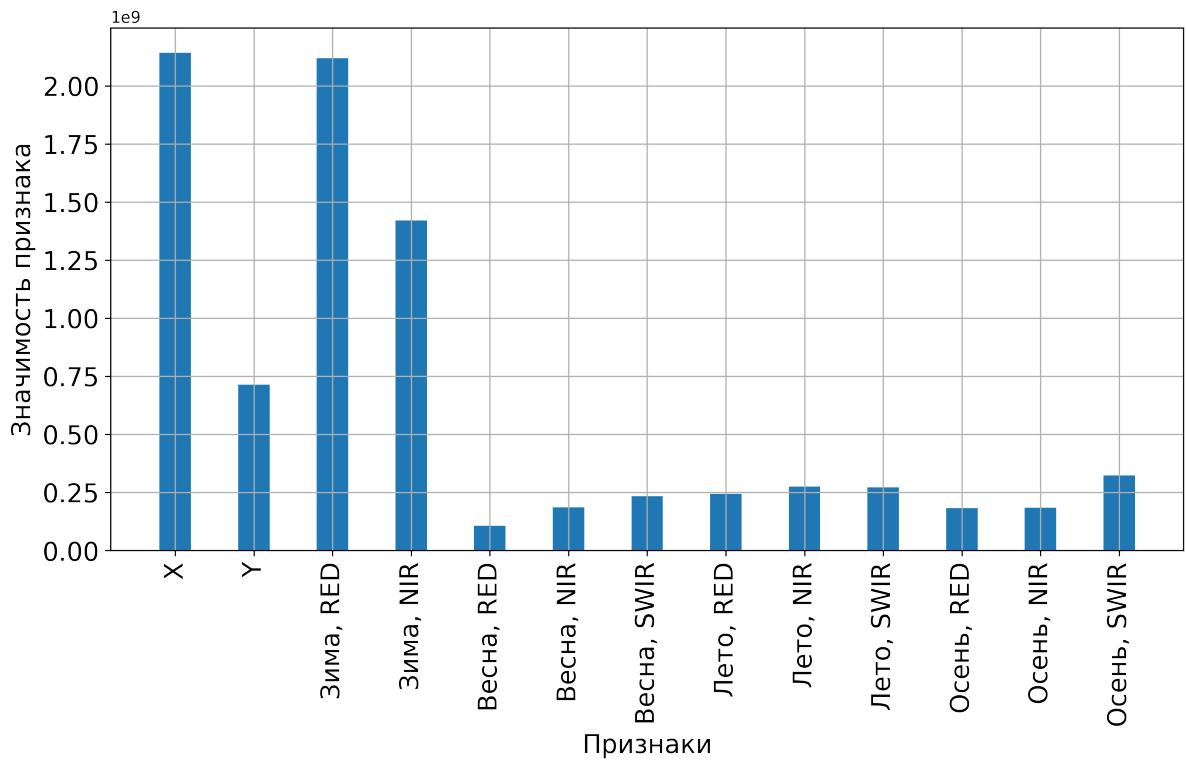
SUMMER1, SUMMER3, WINTER2, FALL3, SPRING3, FALL2, Y, SPRING2, FALL1, SPRING1.

4.2 Оценка значимости признаков с помощью теста

$$\chi^2$$

Тест χ^2 используется в статистике для проверки независимости двух событий. Поскольку нам необходимо отобрать признаки, наиболее зависимые от метки класса, вычислим χ^2 между каждым признаком и меткой класса. Более значимыми являются признаки с наибольшими значениями.

Рис. 10: Оценка значимости признаков с помощью теста χ^2

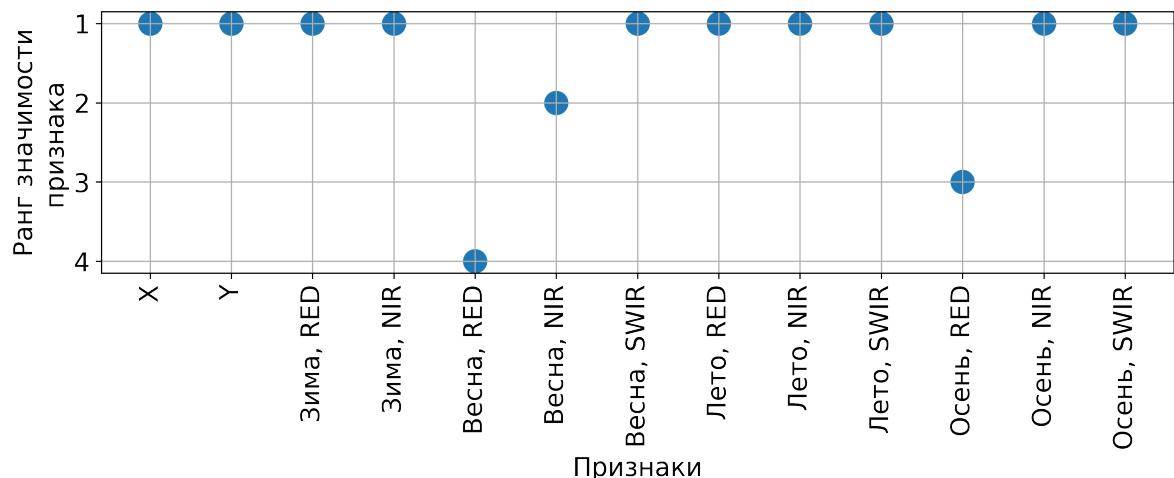


Согласно оценке значимости признаков с помощью теста χ^2 , отсортированная по убыванию значимости последовательность признаков выглядит следующим образом: X, WINTER1, WINTER2, Y, FALL3, SUMMER2, SUMMER3, SUMMER1, SPRING3, SPRING2, FALL2, FALL1, SPRING1.

4.3 Оценка значимости признаков с помощью рекурсивного исключения

Оценка значимости признаков с помощью рекурсивного исключения осуществляется путем рассмотрения все меньших и меньших наборов признаков. Сначала оценочная система обучается на начальном наборе признаков, и важность каждого признака определяется либо через какой-либо конкретный атрибут, либо через вызываемый признак. Затем из текущего набора признаков отсекаются наименее важные. Эта процедура рекурсивно повторяется на обрезанном наборе до тех пор, пока не будет достигнуто желаемое количество выбираемых признаков.

Рис. 11: Оценка значимости признаков с помощью рекурсивного исключения (более значимые признаки находятся на графике выше менее значимых)



Согласно оценке значимости признаков с помощью рекурсивного исключения, отсортированная по убыванию значимости последовательность признаков выглядит следующим образом: X, Y, WINTER1, WINTER2, SPRING3, SUMMER1, SUMMER2, SUMMER3, FALL2, FALL3, SPRING2, FALL1, SPRING1.

4.4 Обучение случайного леса и оценка качества классификации на выборке, содержащей наиболее значимые признаки

Произведено обучение случайного леса с гиперпараметрами по умолчанию на обучающей выборке (200000 элементов), содержащей наиболее значимые признаки и получены предсказания типов земного покрова для тестовой выборки (1800000 элементов), также содержащей наиболее значимые признаки. Ниже приведены оценки качества классификации тестовой выборки обученной моделью случайного леса:

$$Accuracy = 0.95441,$$

$$F_1\text{-score} = 0.95365 \text{ (при взвешенном усреднении)},$$

$$Out-of-bag error = 0.95328.$$

Рис. 12: Матрица ошибок классификации случайного леса, обученного на выборке, содержащей наиболее значимые признаки

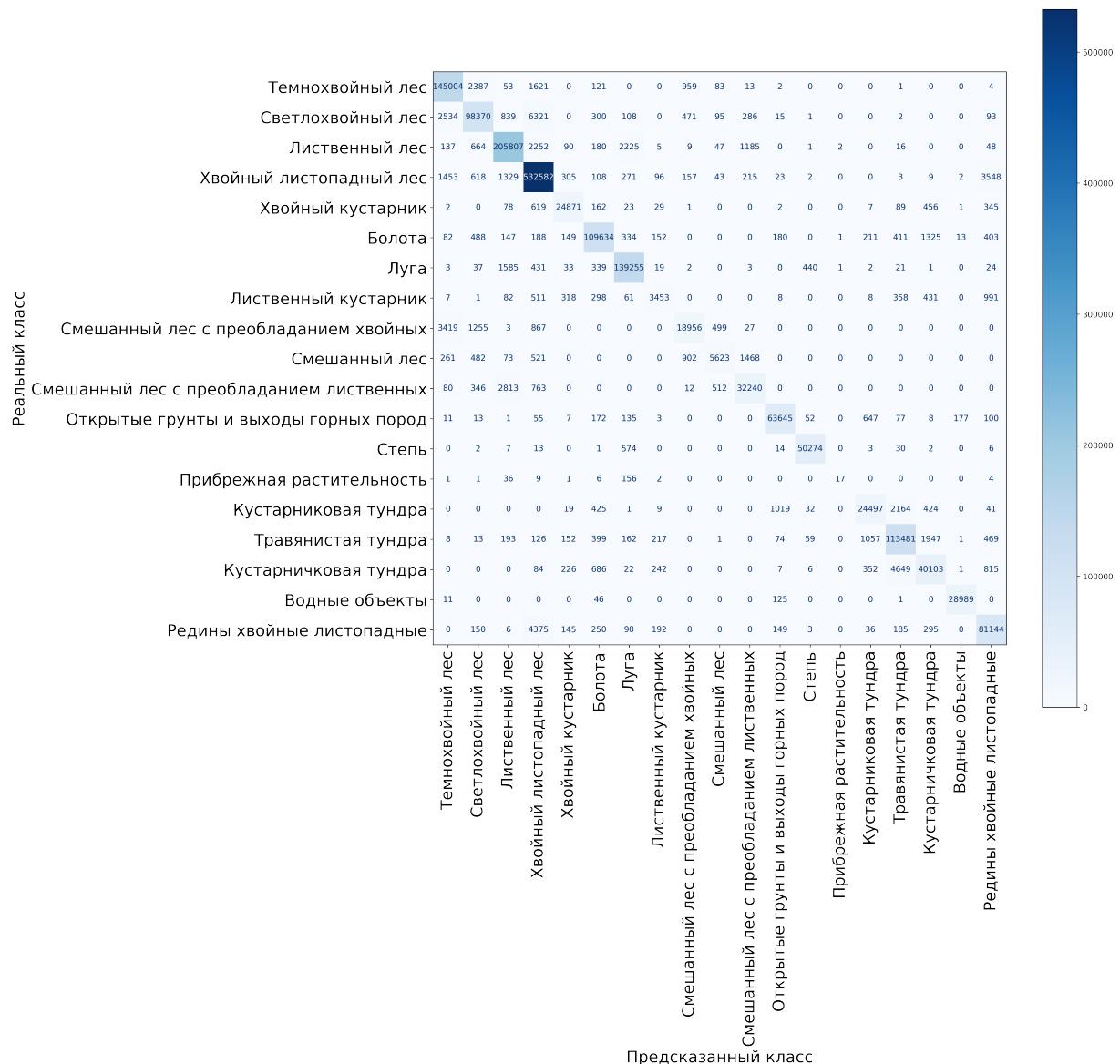
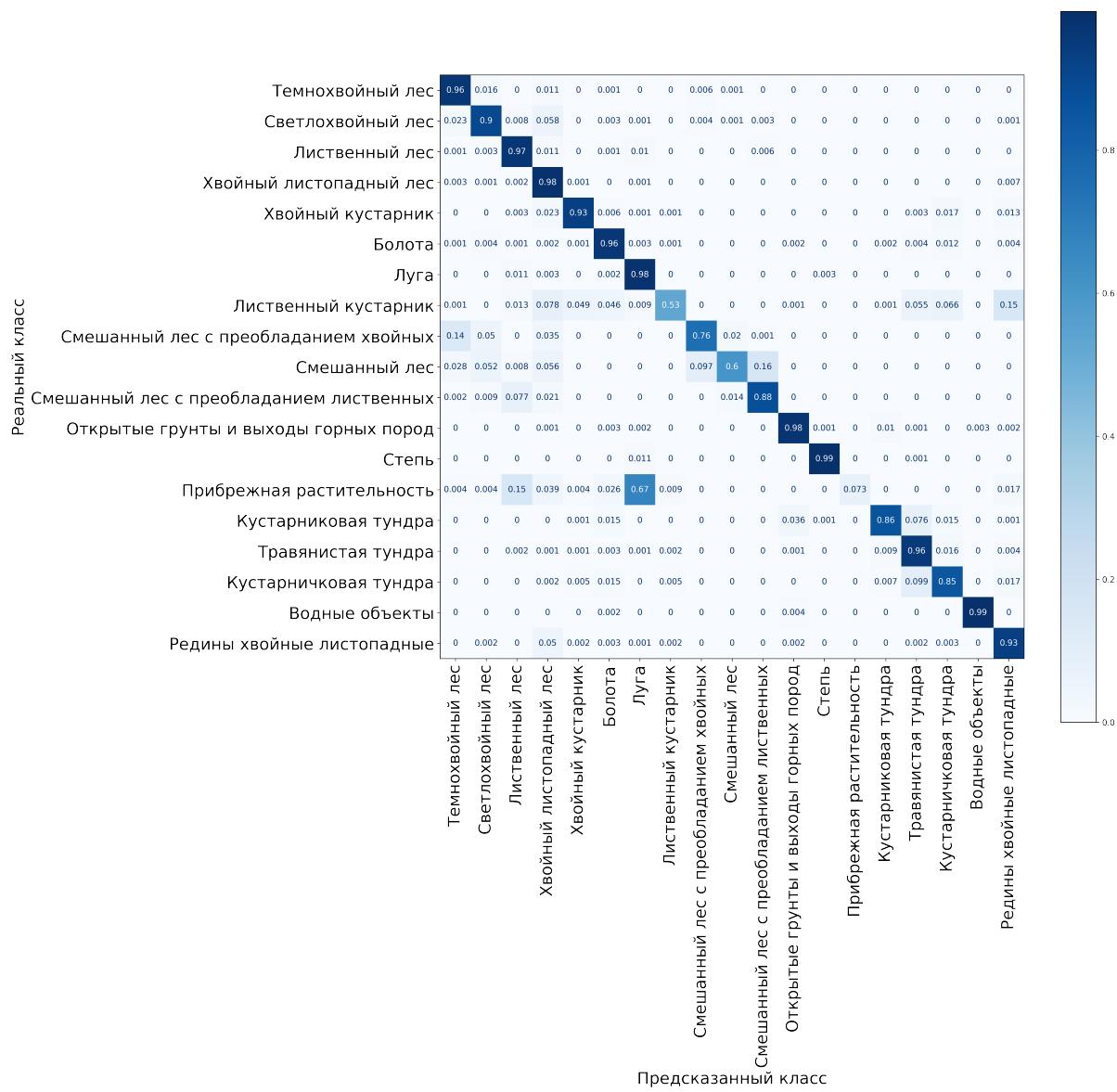


Рис. 13: Нормированная по строкам матрица неточностей классификации случайного леса, обученного на выборке, содержащей наиболее значимые признаки



В результате отбора наиболее значимых признаков, мы видим несущественное улучшение качества классификации и уменьшение количества времени, затрачиваемого на обучение модели, примерно на 15%.

5 Заключение

6 Литература

- [1] Барталев С.А., Егоров В.А., Жарко В.О., Лупян Е.А., Плотников Д.Е., Хвостиков С.А., Шабанов Н.В. Спутниковое картографирование растительного покрова России // ИКИ РАН. 2016. С. 93-110.
- [2] Барталев С.А., Егоров В.А., Жарко В.О., Лупян Е.А., Плотников Д.Е., Хвостиков С.А. Состояние и перспективы развития методов спутникового картографирования растительного покрова России // Современные проблемы дистанционного зондирования Земли из космоса. 2015. Т. 12. № 5. С. 203-221.
- [3] Belgiu M., Drăguț L. Random forest in remote sensing: A review of applications and future directions // ISPRS Journal of Photogrammetry and Remote Sensing. 2016. No. 114. pp. 24-31.
- [4] Документация языка Python
<https://docs.python.org/3/>
- [5] Документация библиотеки Pandas
<https://pandas.pydata.org/docs/>
- [6] Документация библиотеки Vaex
<https://vaex.io/docs/>
- [7] Документация библиотеки Matplotlib
<https://matplotlib.org/stable/>
- [8] Документация библиотеки Scikit-learn
<https://scikit-learn.org/stable/>

7 Приложения

↗ Репозиторий проекта на GitHub

<https://github.com/eugeuie/masters-term-paper>