



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

TODO Аннотация на английском + презентация для английского

Abstract

Содержание

1 Введение	4
2 Анализ данных	5
2.1 Источник данных	5
2.2 Общие сведения о данных	6
2.3 Классифицируемые типы земного покрова	7
2.3.1 Тематические классы земного покрова и их пред- ставленность в выборке	7
2.3.2 Визуализация тематических классов земного покрова	11
2.4 Инструменты для анализа данных	34
3 Классификация земного покрова с помощью модели слу- чайного леса	35
4 Значимость признаков	36
5 Применение модели случайного леса к полному набору данных	37
6 Применение модели случайного леса к карте земного по- крова	38
7 Заключение	39
8 Литература	40
9 Приложения	41

1 Введение

Смысл

Зачем классифицировать земной покров?

Зачем новые методы автоматизации классификации земного покрова?

Об алгоритме LAGMA

Что-то о машинном обучении в применении к колоссальным объемам спутниковых данных

Ссылки на [2] и [3]

Ранее к этому набору данных случайные леса не применялись все-рьез. Соответственно развитие заключается в применении нового метода классификации, оценке влияния разных параметров/подходов на точность классификации, а также в анализе входных данных, их значимости.

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, геопривязанные и скорректированные на атмосферу. Данные предварительно обработаны для исключения влияния облаков и теней от них путем осреднения значений яркости за сезон, по ним созданы сезонные композиты.

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 [1].

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 земного покрова, а именно:

1. Леса:

- Темнохвойные вечнозеленые насаждения (*темнохвойный лес*), в пологе которых не менее 80% площади крон составляют теневыносливые виды хвойных деревьев, включая ель, пихту и сибирскую сосну (кедр).
- Светлохвойные вечнозеленые насаждения (*светлохвойный лес*), в пологе которых не менее 80% площади крон составляют 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 площади 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



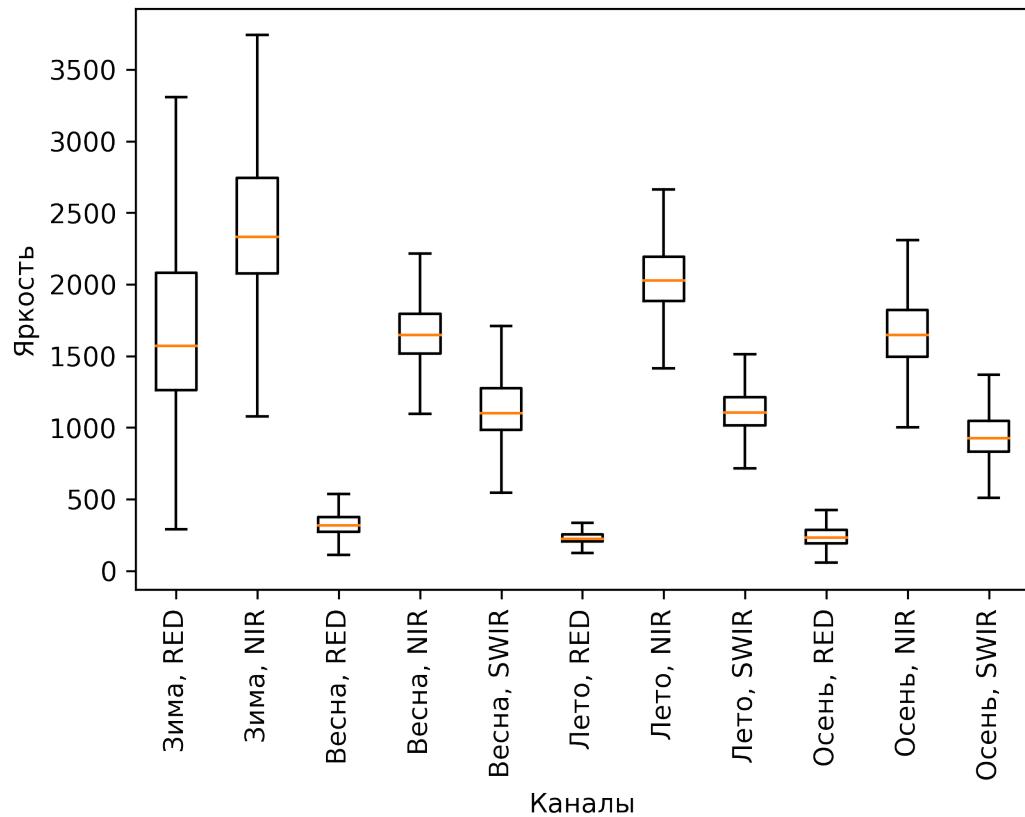
2.3.2 Визуализация тематических классов земного покрова

TODO мотивация визуализации, выводы какие-то о том, что признаки у всех типов разные и можно использовать случайный лес? TODO мотивация визуализации, выводы какие-то о том, что признаки у всех типов разные и можно использовать случайный лес? TODO мотивация

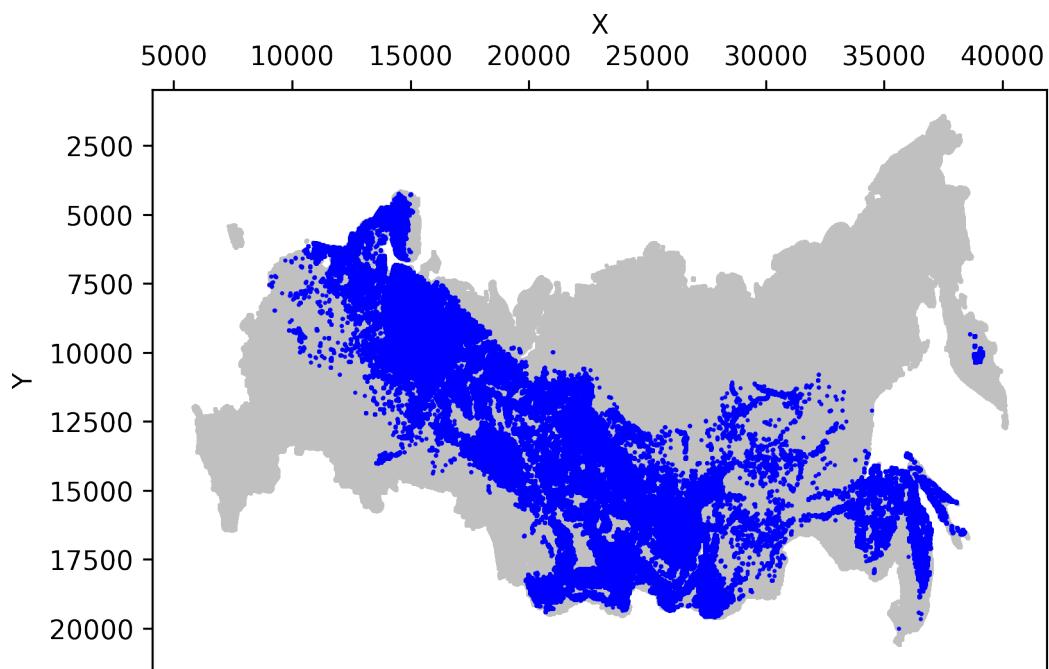
визуализации, выводы какие-то о том, что признаки у всех типов разные и можно использовать случайный лес? TODO мотивация визуализации, выводы какие-то о том, что признаки у всех типов разные и можно использовать случайный лес? TODO мотивация визуализации, выводы какие-то о том, что признаки у всех типов разные и можно использовать случайный лес? TODO мотивация визуализации, выводы какие-то о том, что признаки у всех типов разные и можно использовать случайный лес? TODO мотивация визуализации, выводы какие-то о том, что признаки у всех типов разные и можно использовать случайный лес? TODO мотивация визуализации, выводы какие-то о том, что признаки у всех типов разные и можно использовать случайный лес? TODO мотивация визуализации, выводы какие-то о том, что признаки у всех типов разные и можно использовать случайный лес? TODO мотивация визуализации, выводы какие-то о том, что признаки у всех типов разные и можно использовать случайный лес? TODO мотивация визуализации, выводы какие-то о том, что признаки у всех типов разные и можно использовать случайный лес? TODO мотивация визуализации, выводы какие-то о том, что признаки у всех типов разные и можно использовать случайный лес? TODO мотивация визуализации, выводы какие-то о том, что признаки у всех типов разные и можно использовать случайный лес? TODO мотивация визуализации, выводы какие-то о том, что признаки у всех типов разные и можно использовать случайный лес? TODO мотивация визуализации, выводы какие-то о том, что признаки у всех типов разные и можно использовать случайный лес? TODO мотивация визуализации, выводы какие-то о том, что признаки у всех типов разные и можно использовать случайный лес?

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).

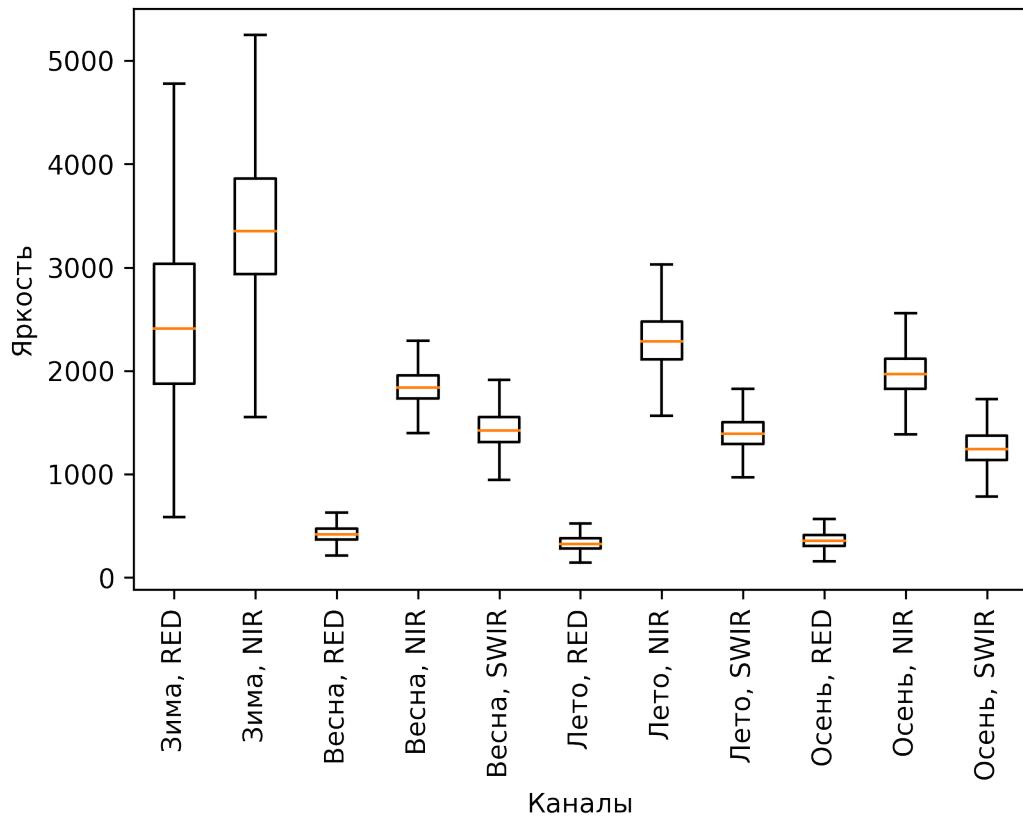
Темнохвойный лес



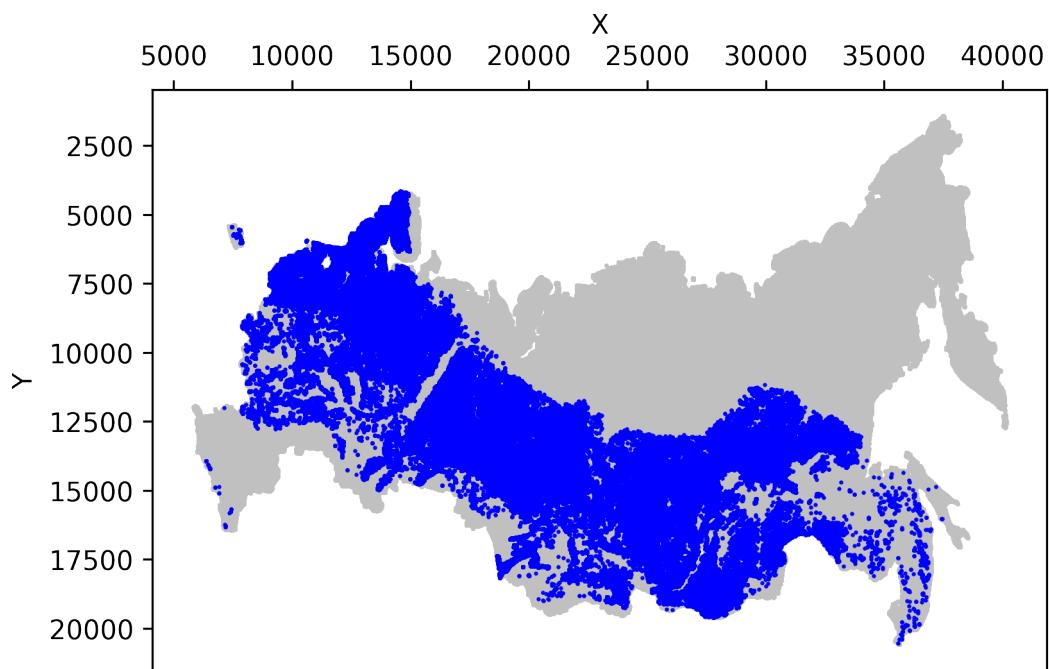
Темнохвойный лес



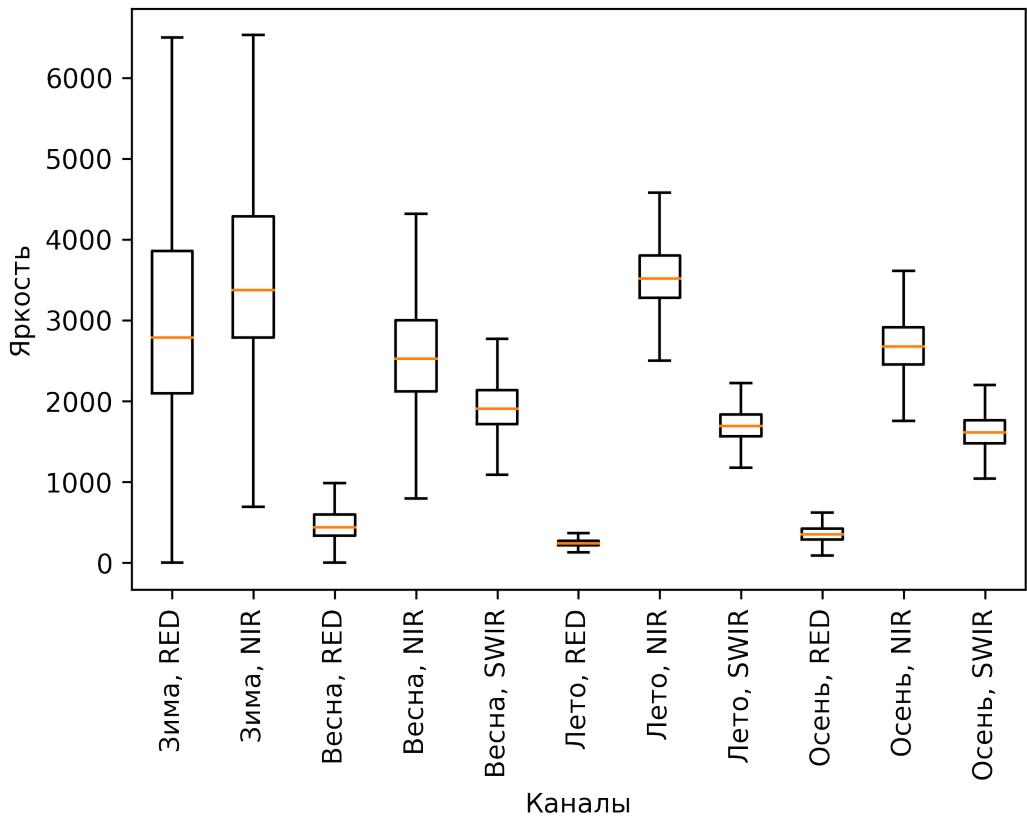
Светлохвойный лес



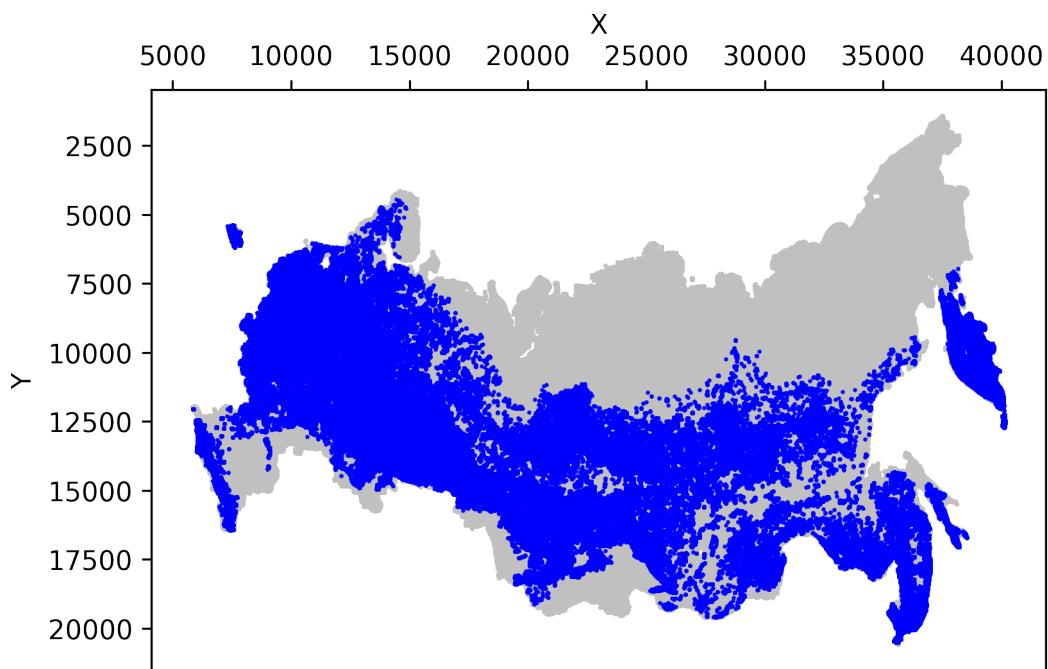
Светлохвойный лес



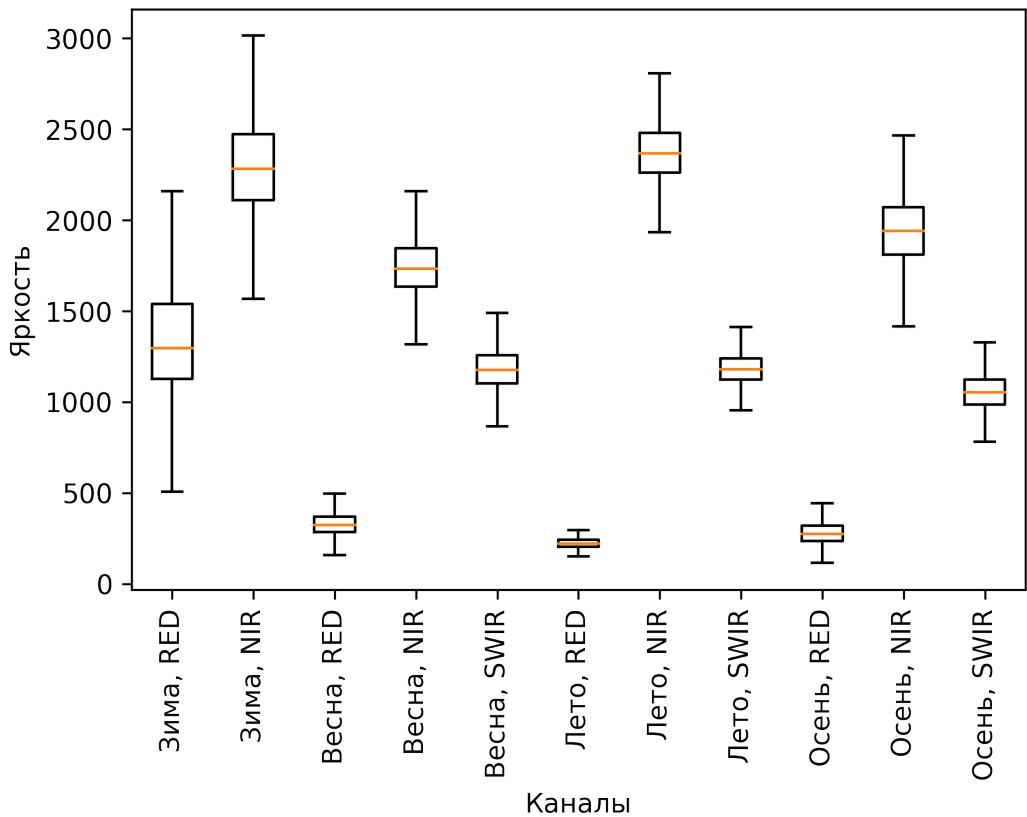
Лиственый лес



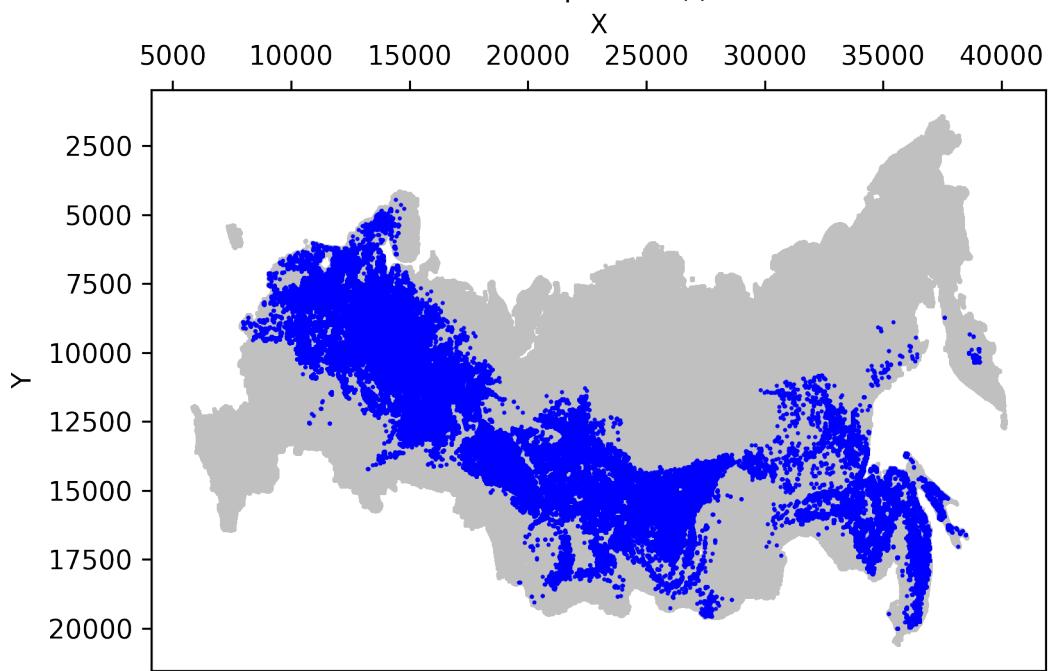
Лиственый лес



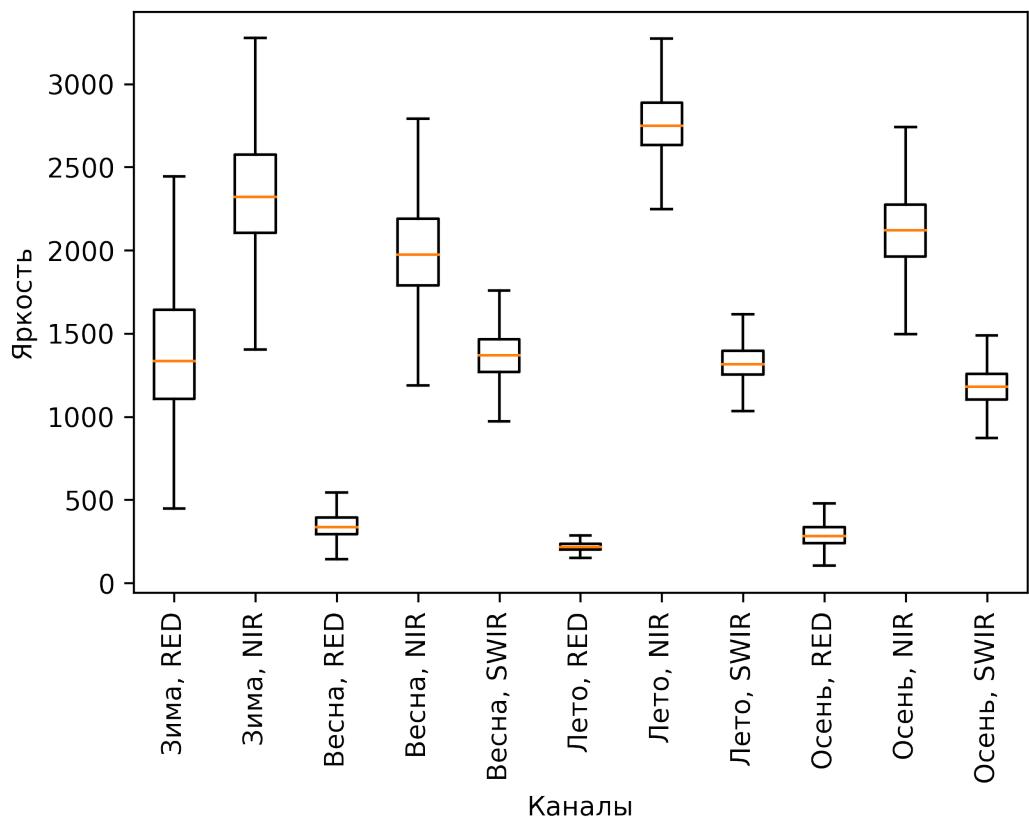
Смешанный лес с преобладанием хвойных



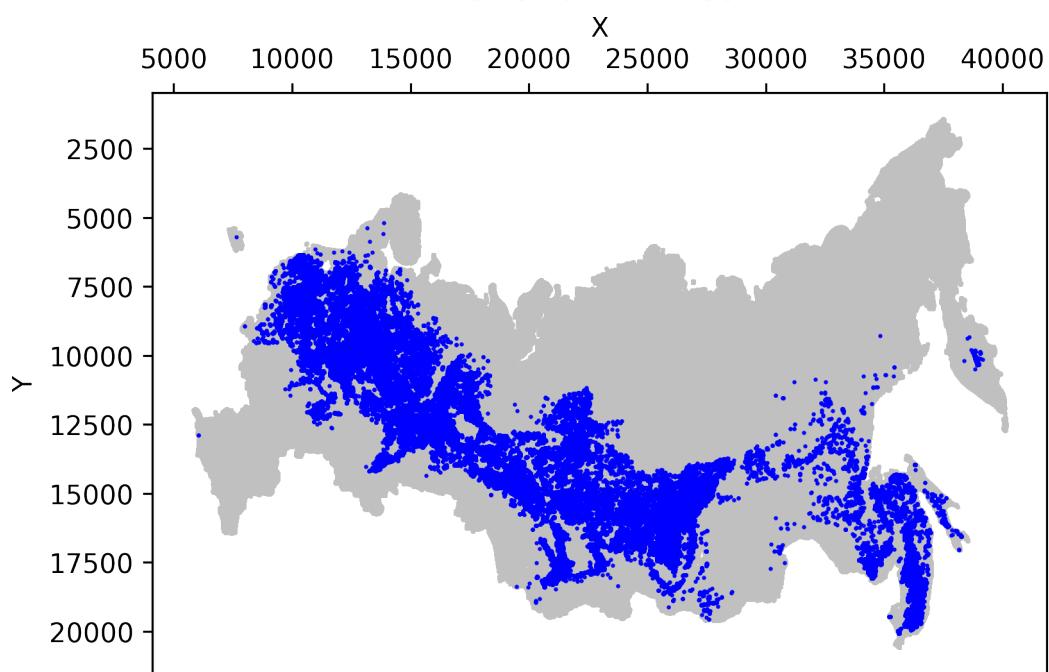
Смешанный лес с преобладанием хвойных



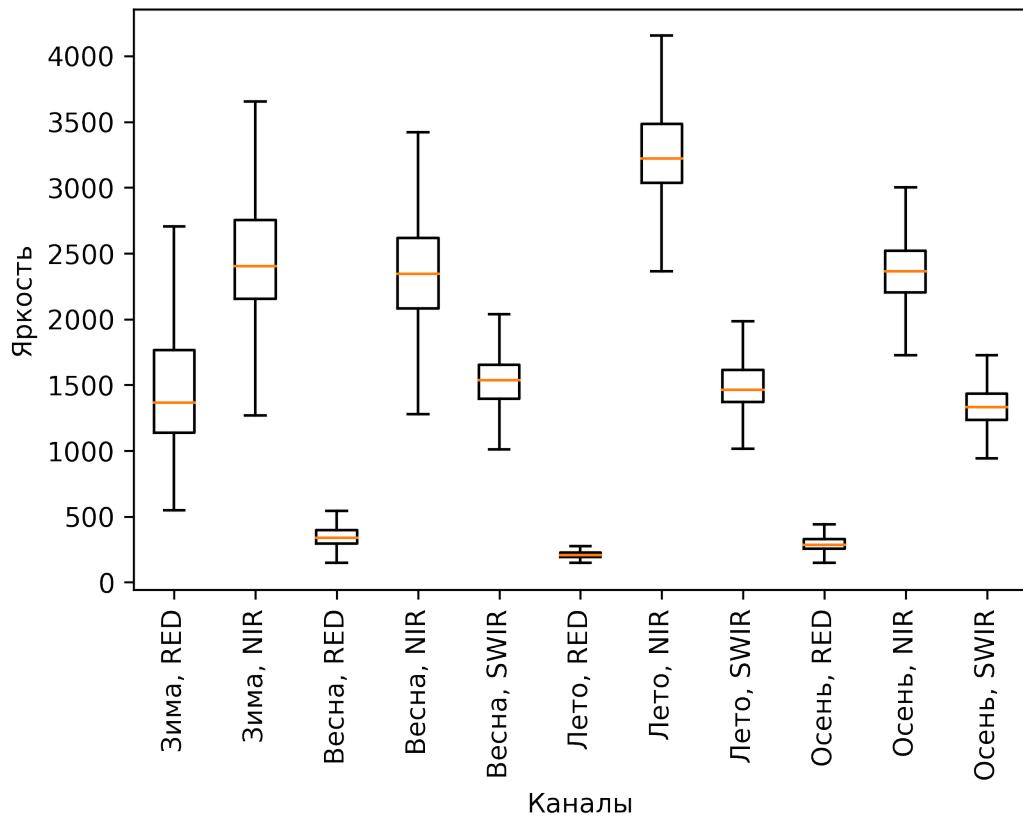
Смешанный лес



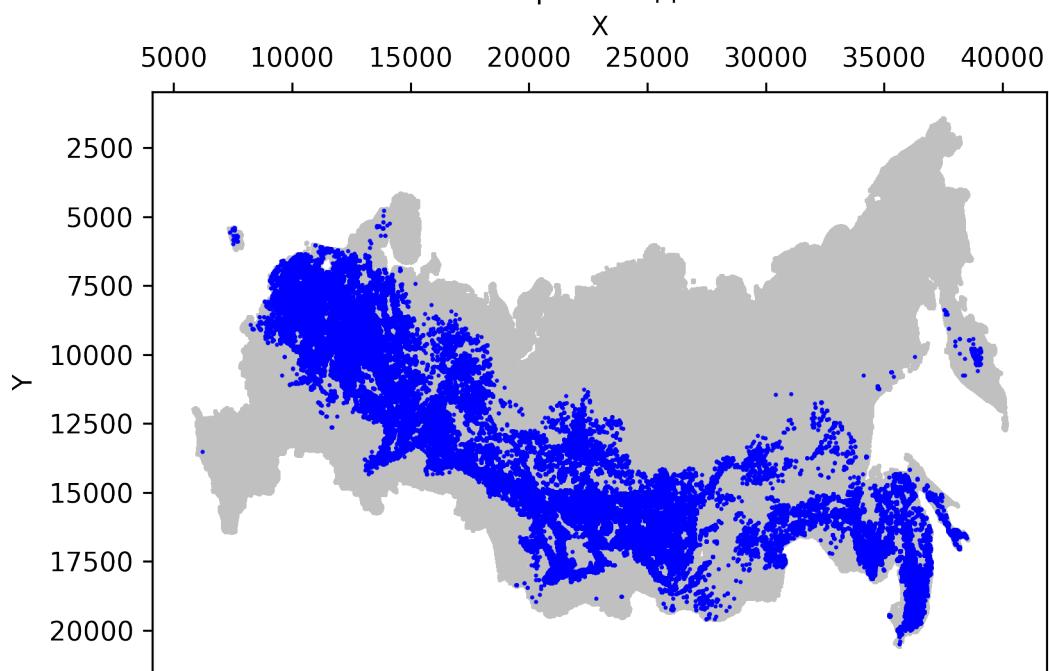
Смешанный лес



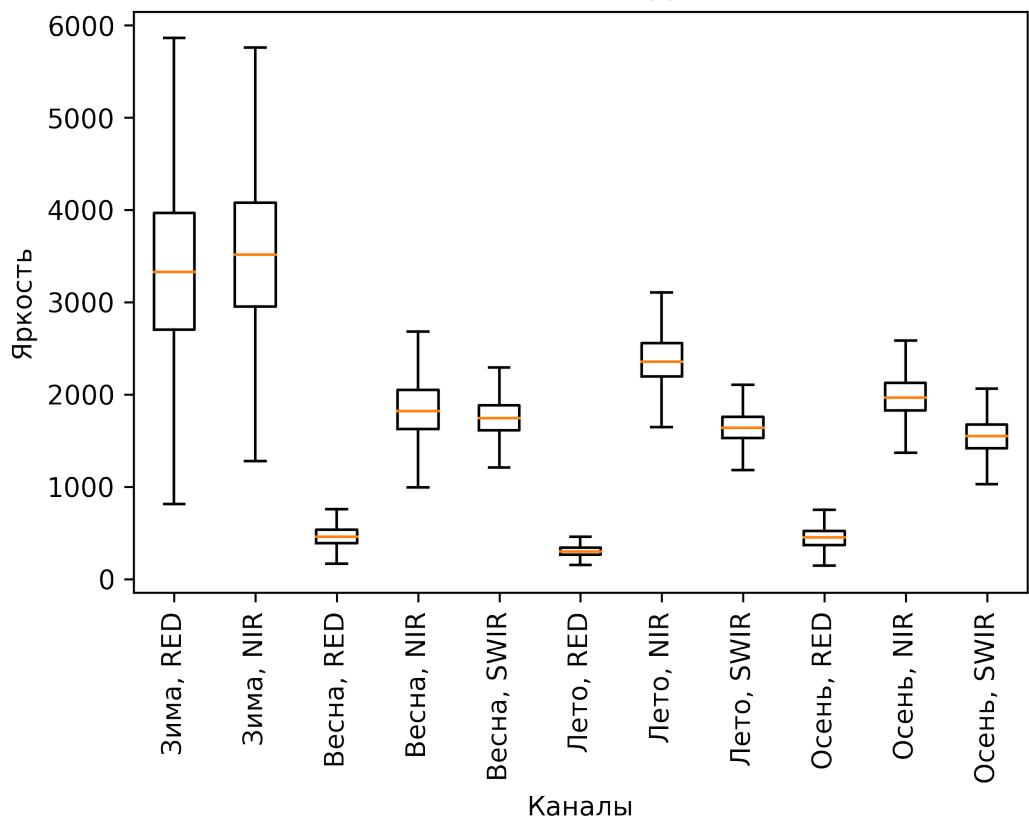
Смешанный лес с преобладанием лиственных



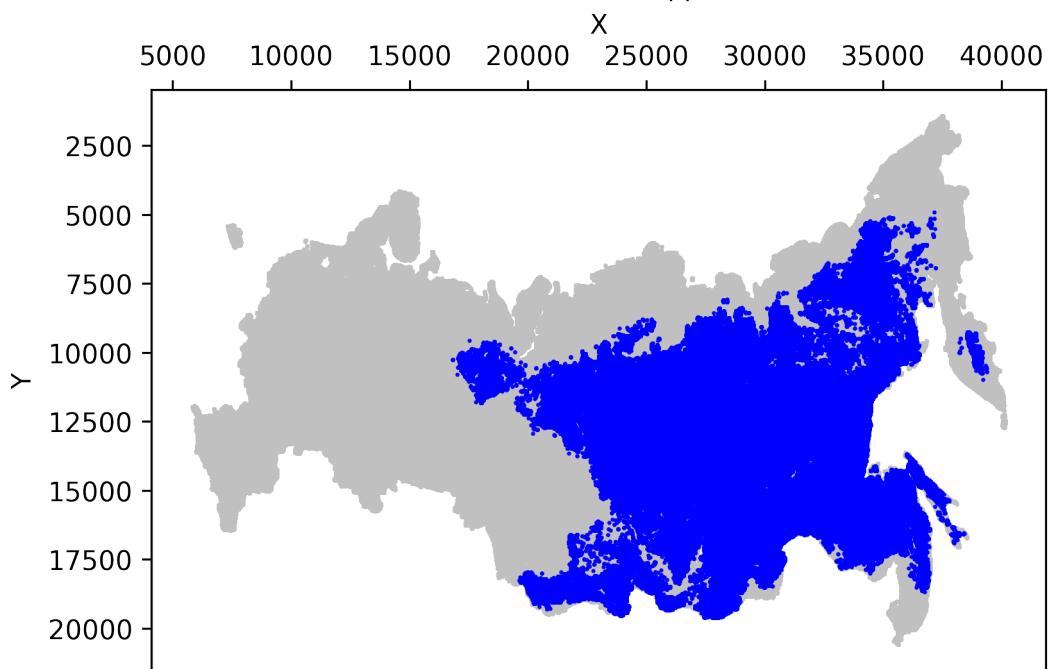
Смешанный лес с преобладанием лиственных



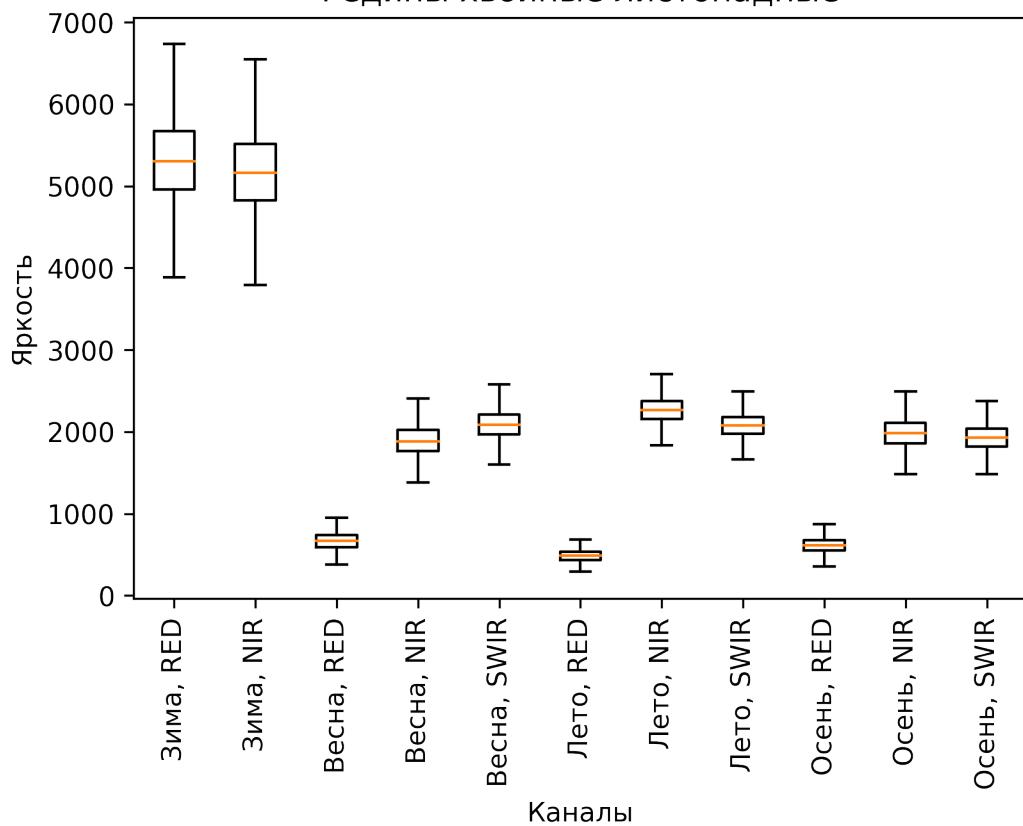
Хвойный листопадный лес



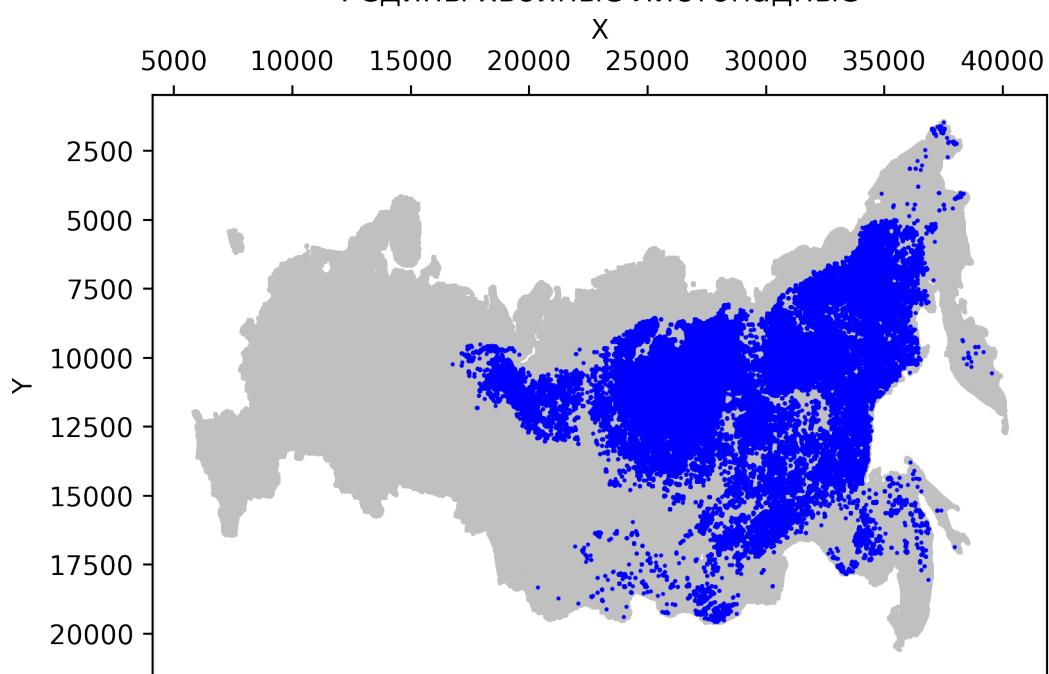
Хвойный листопадный лес



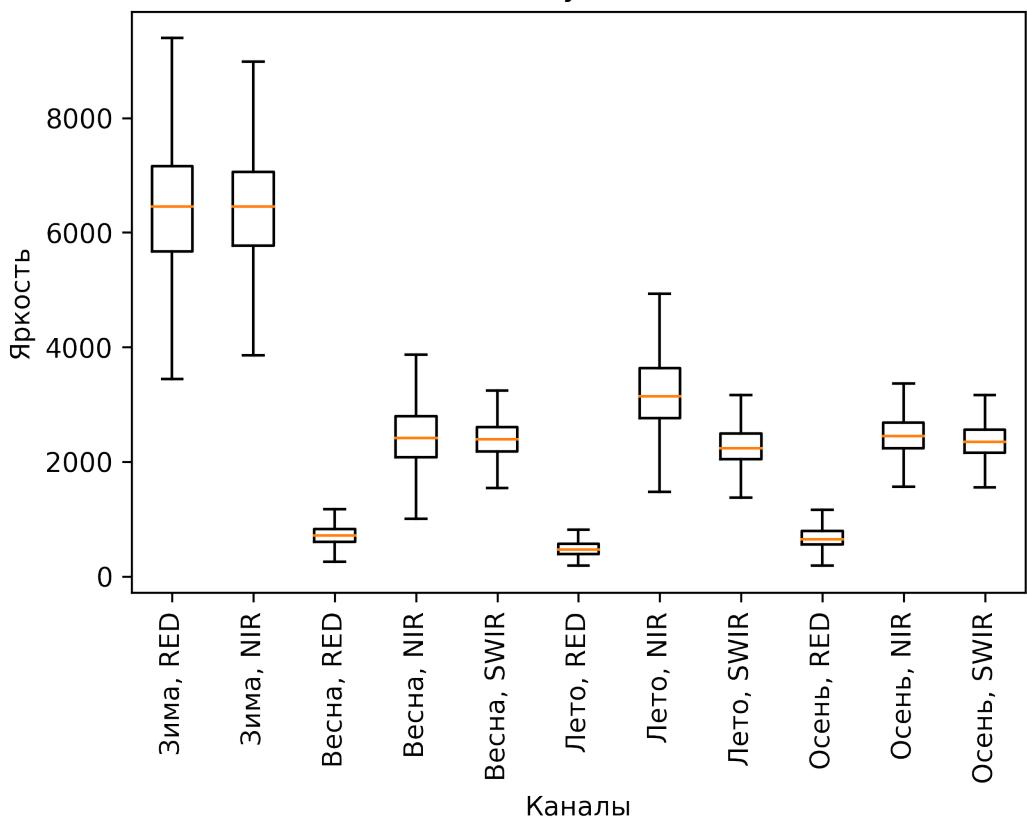
Редины хвойные листопадные



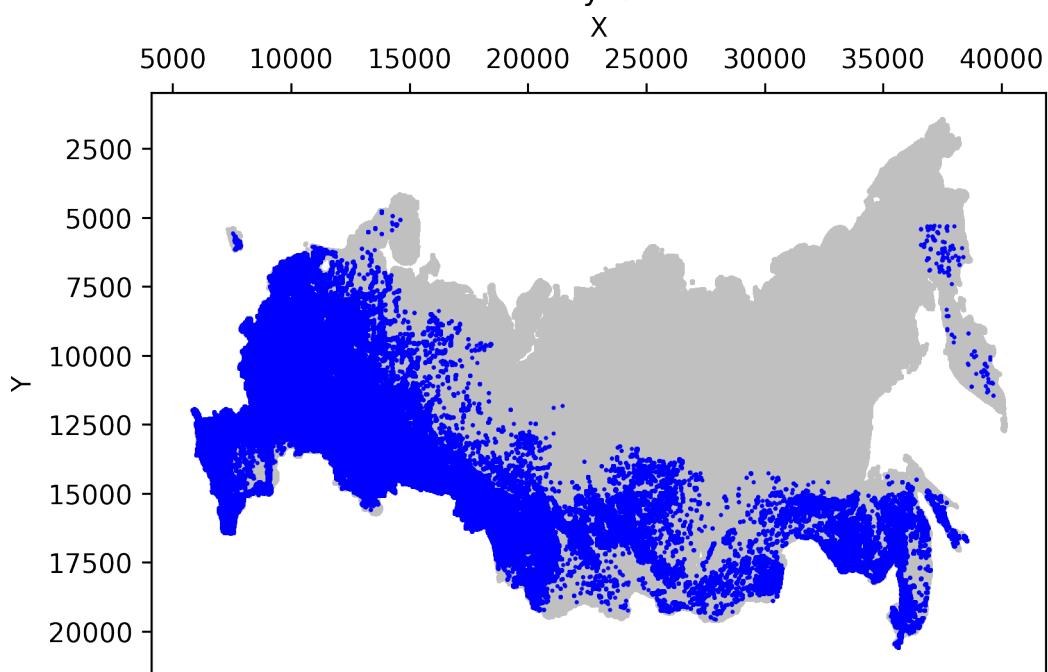
Редины хвойные листопадные

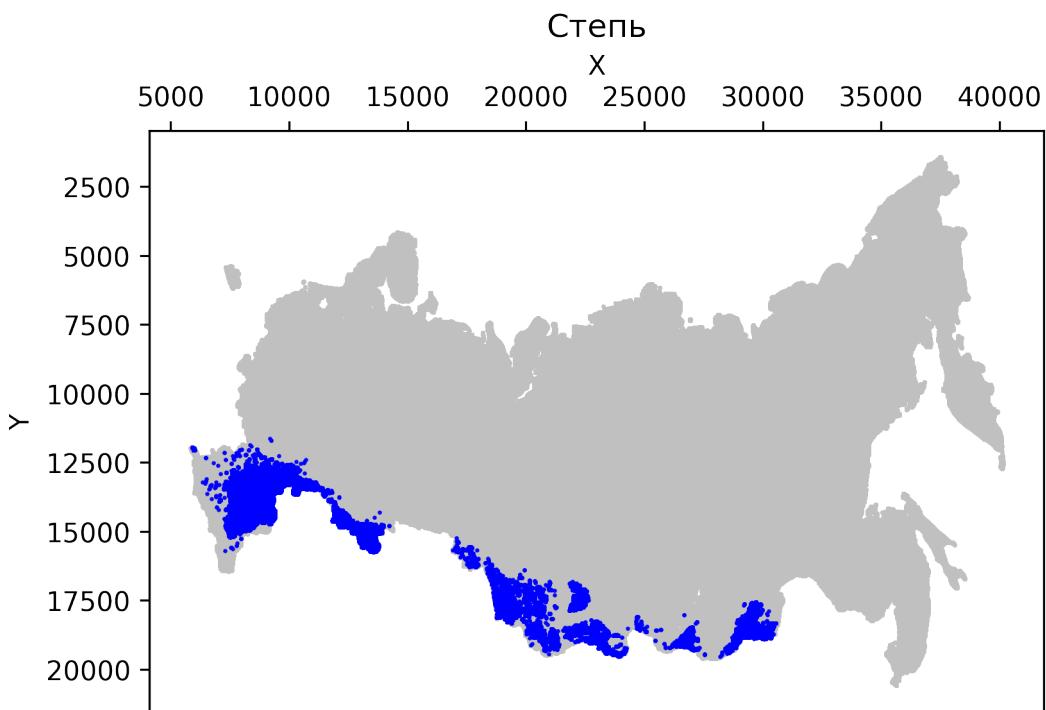
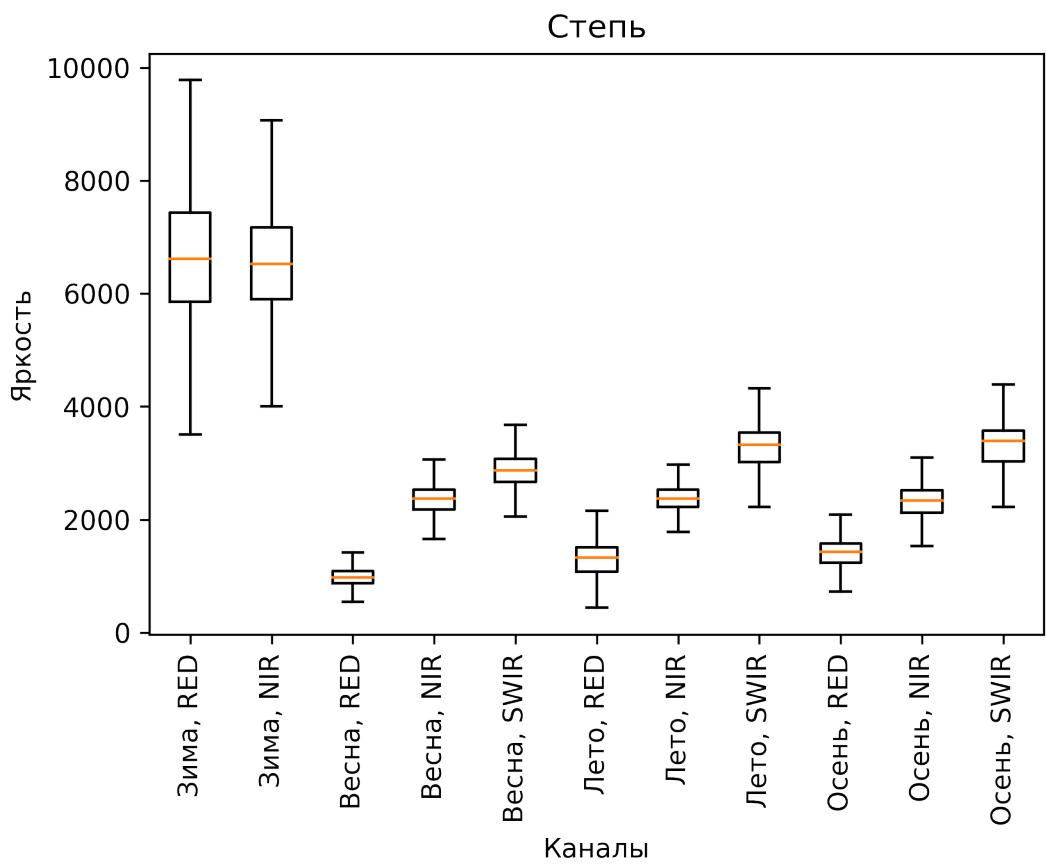


Луга

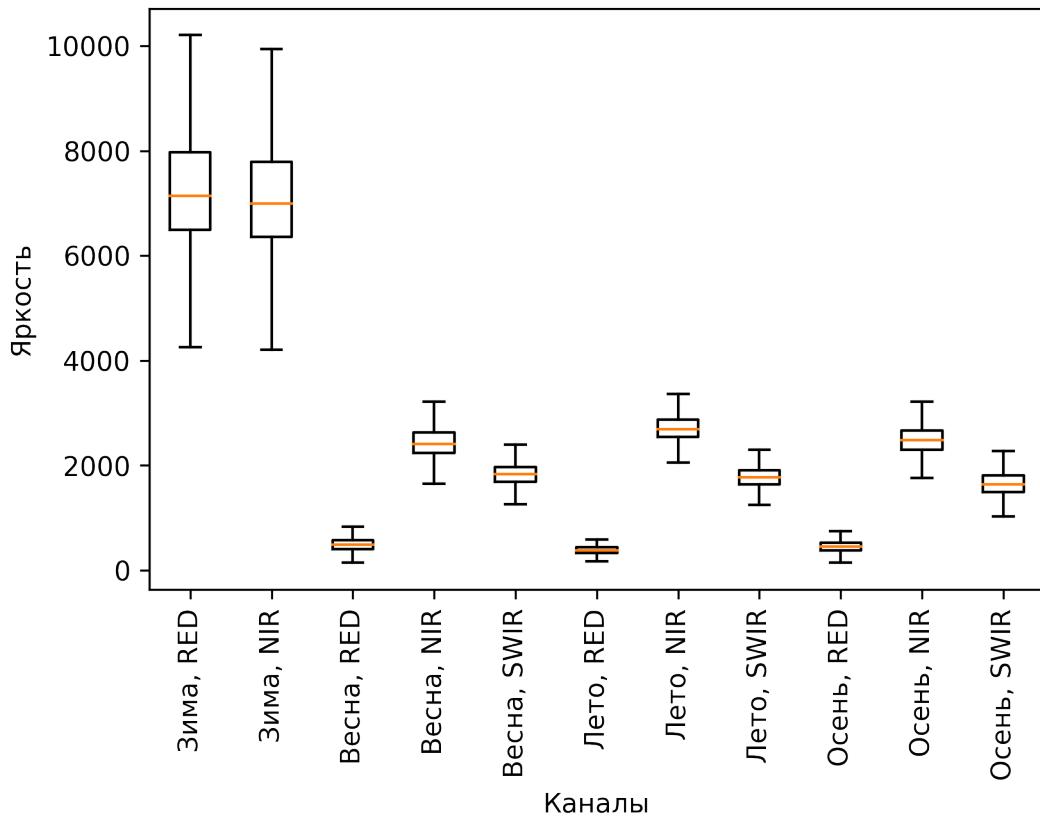


Луга

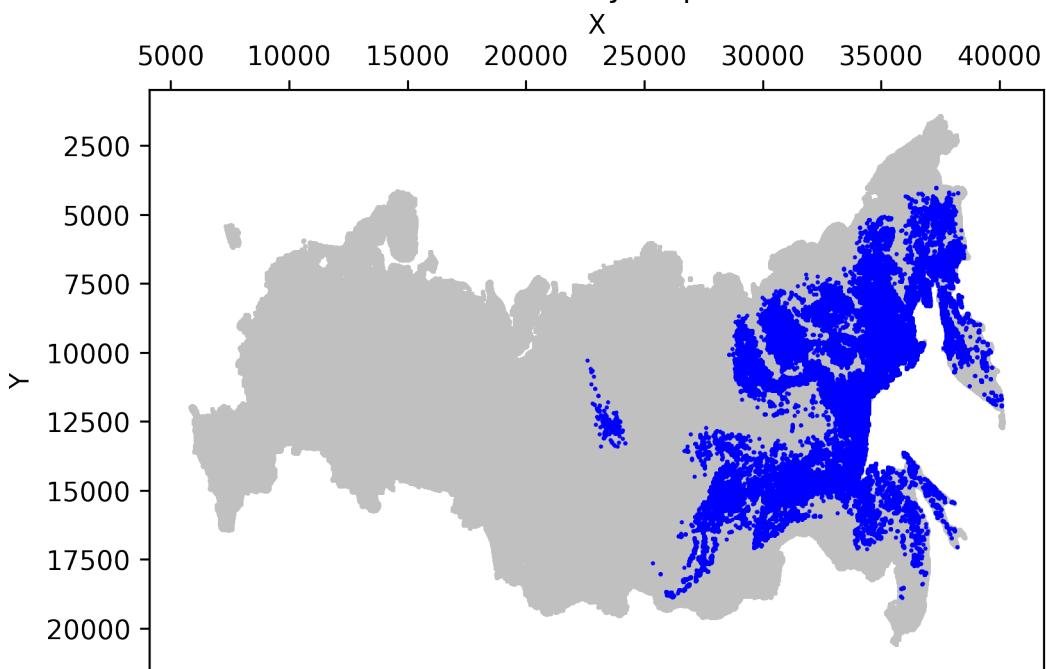




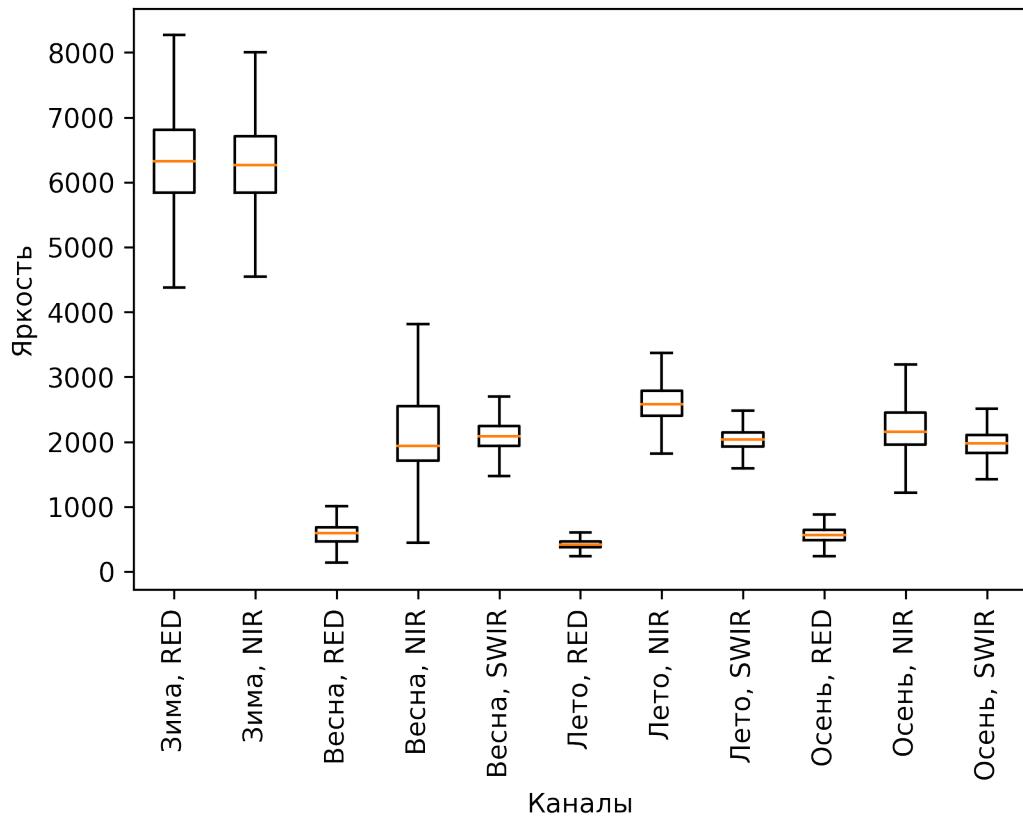
Хвойный кустарник



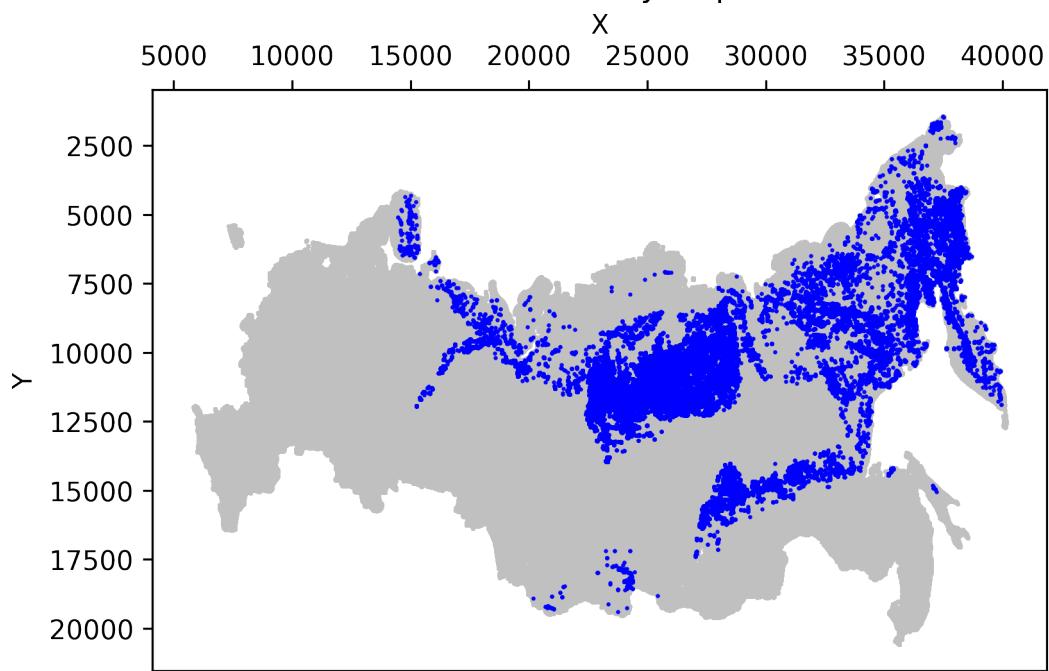
Хвойный кустарник



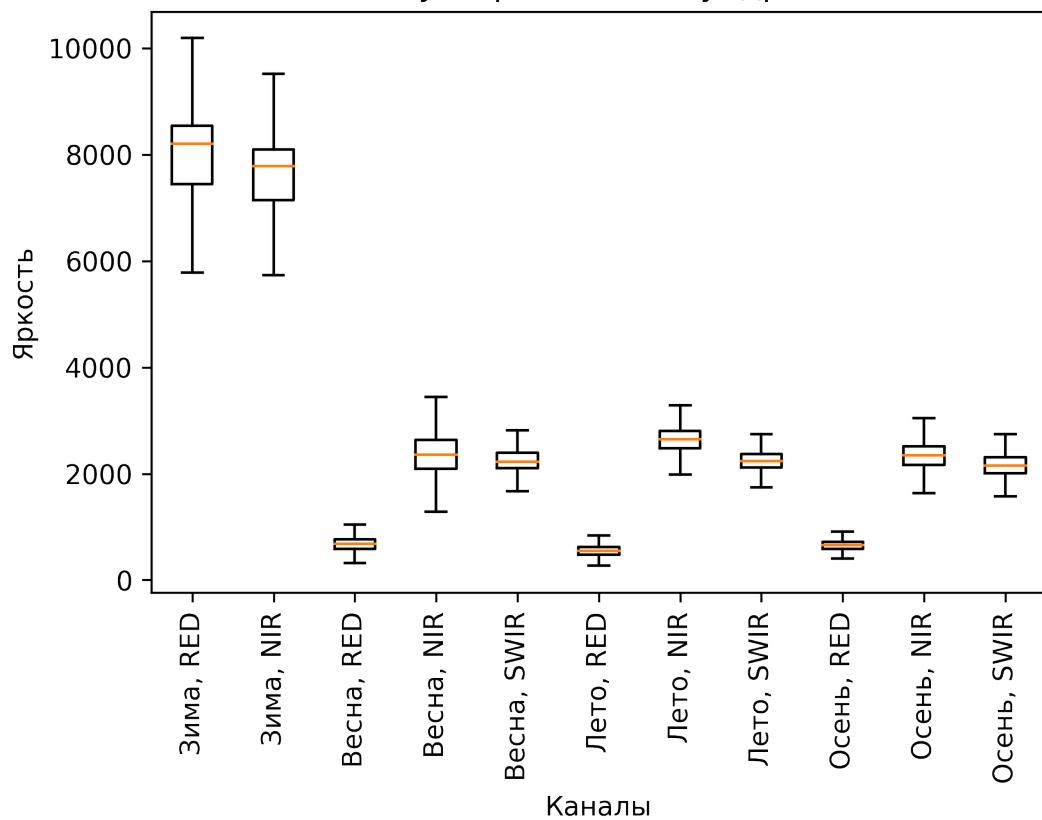
Лиственый кустарник



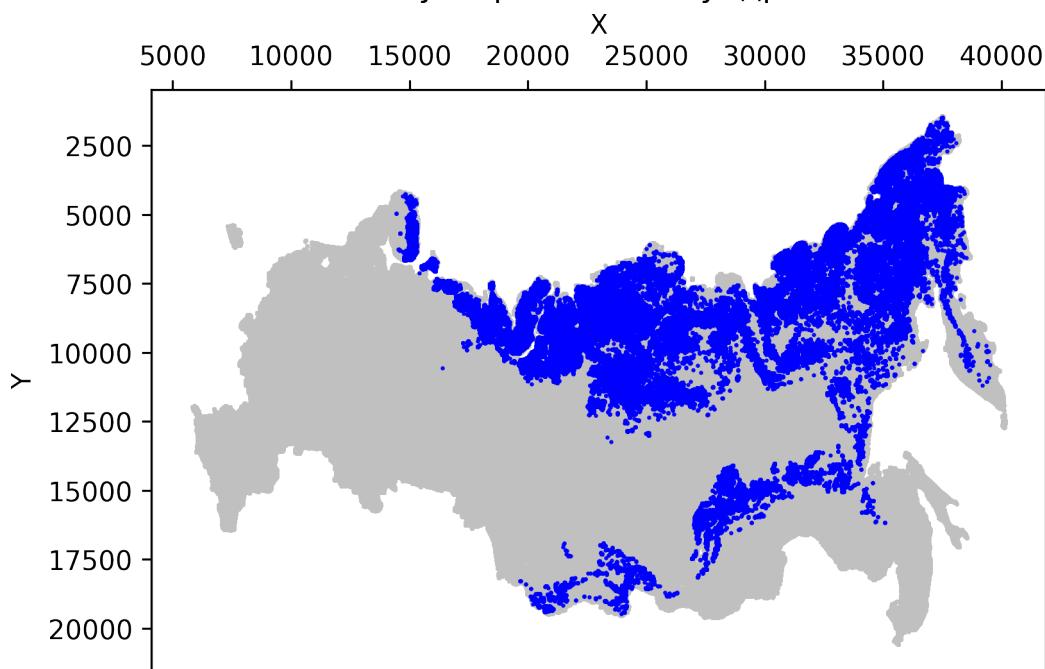
Лиственый кустарник



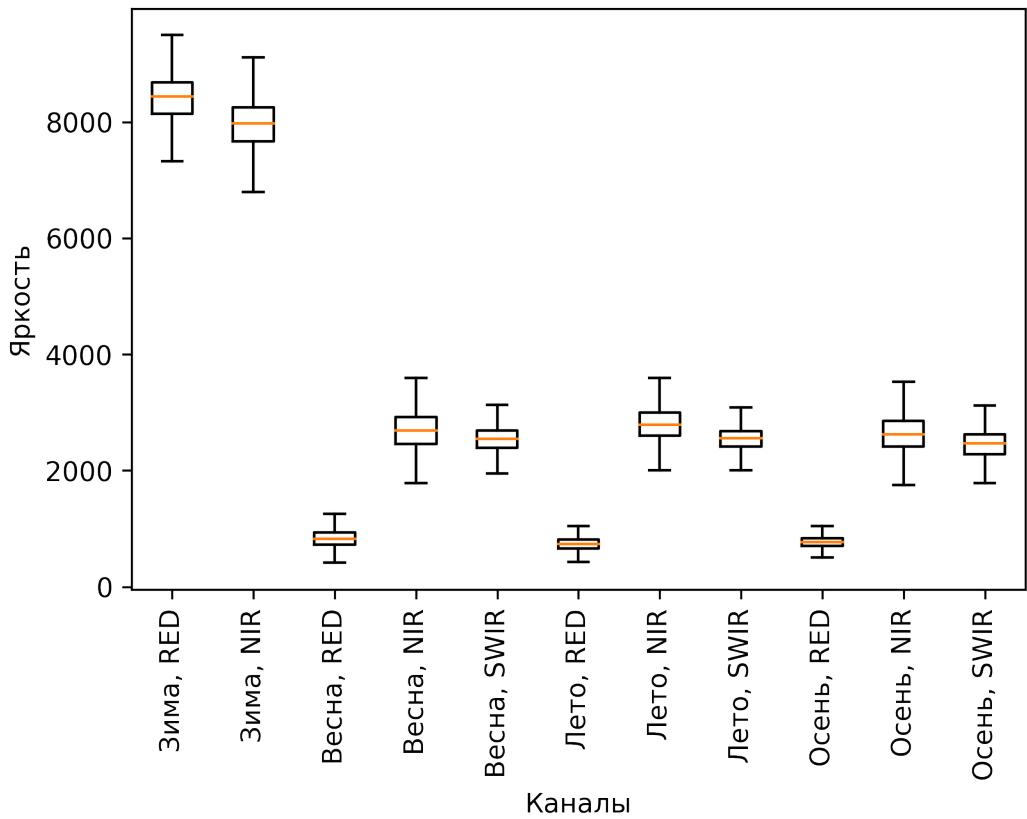
Кустарничковая тундра



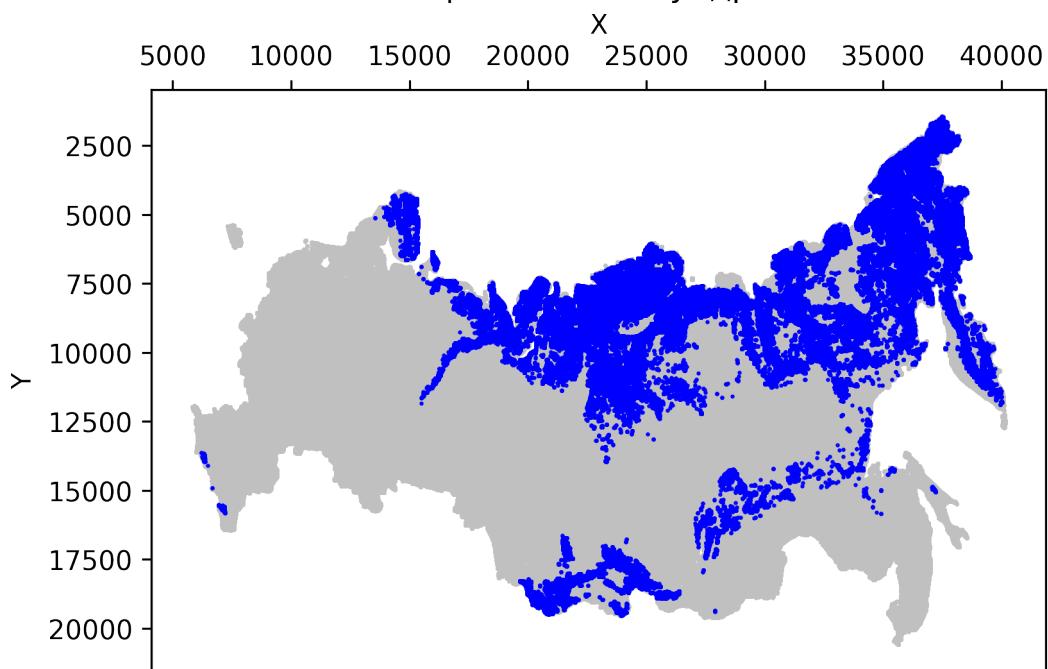
Кустарничковая тундра

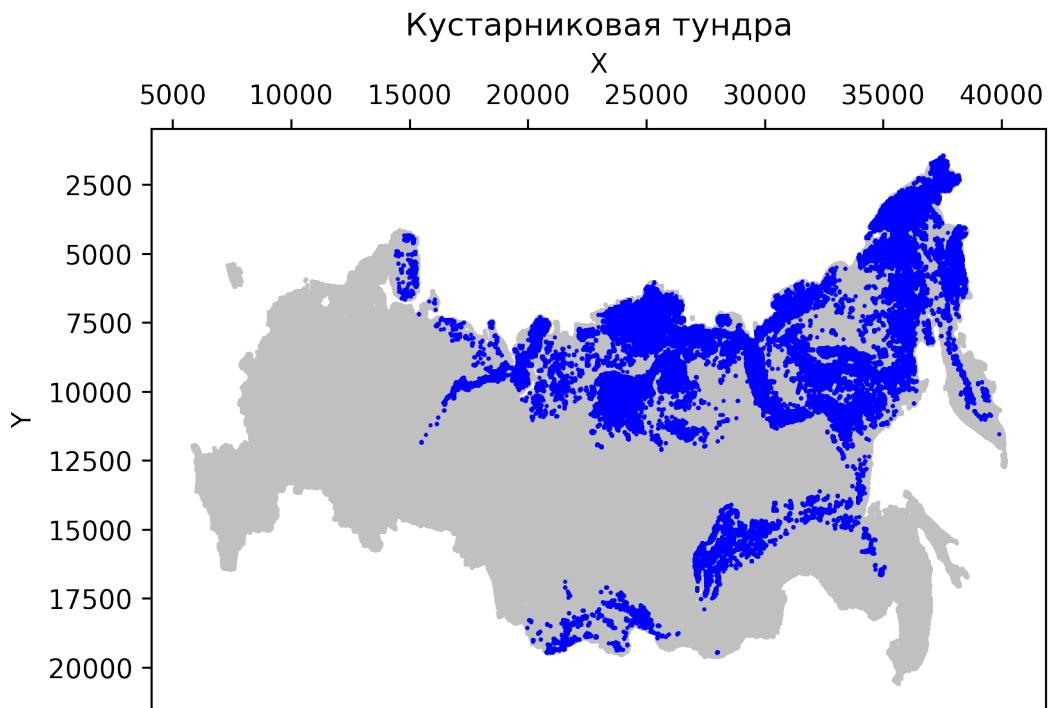
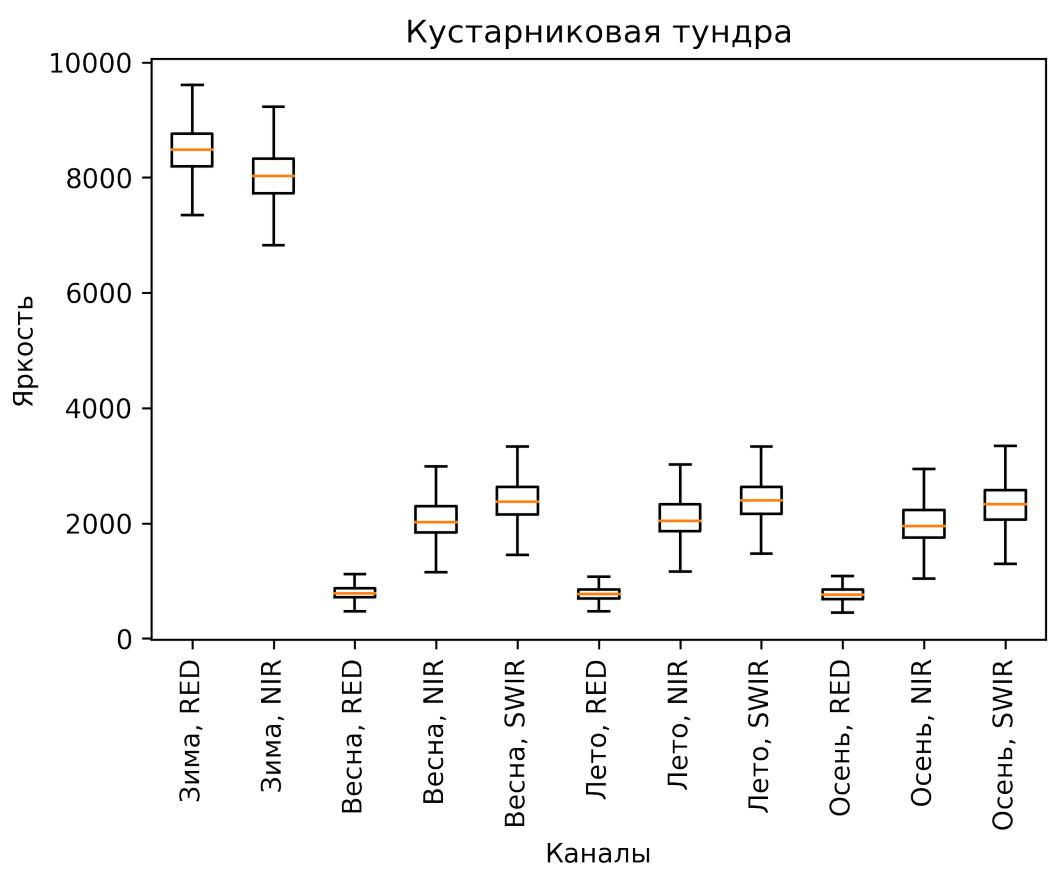


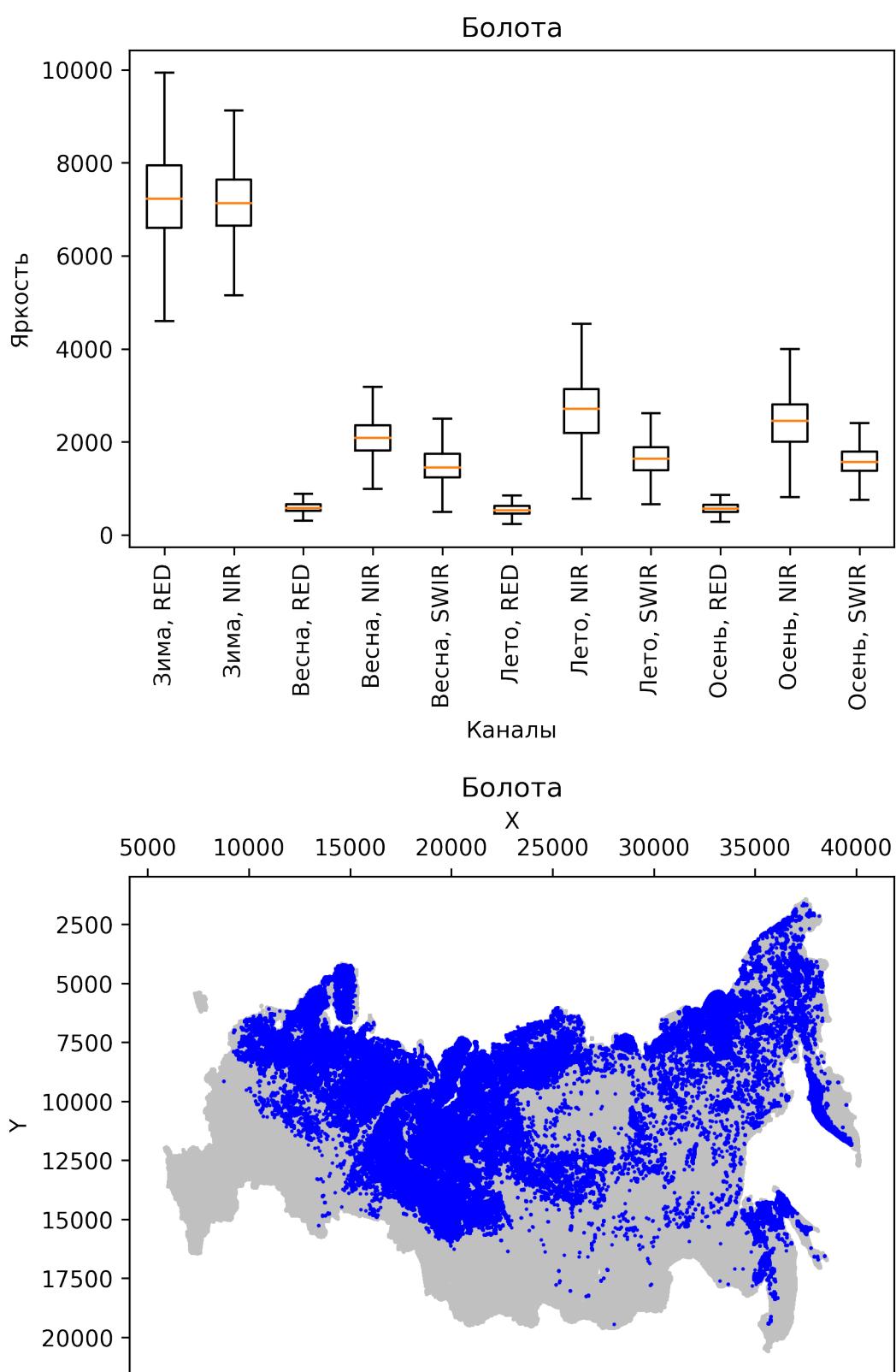
Травянистая тундра



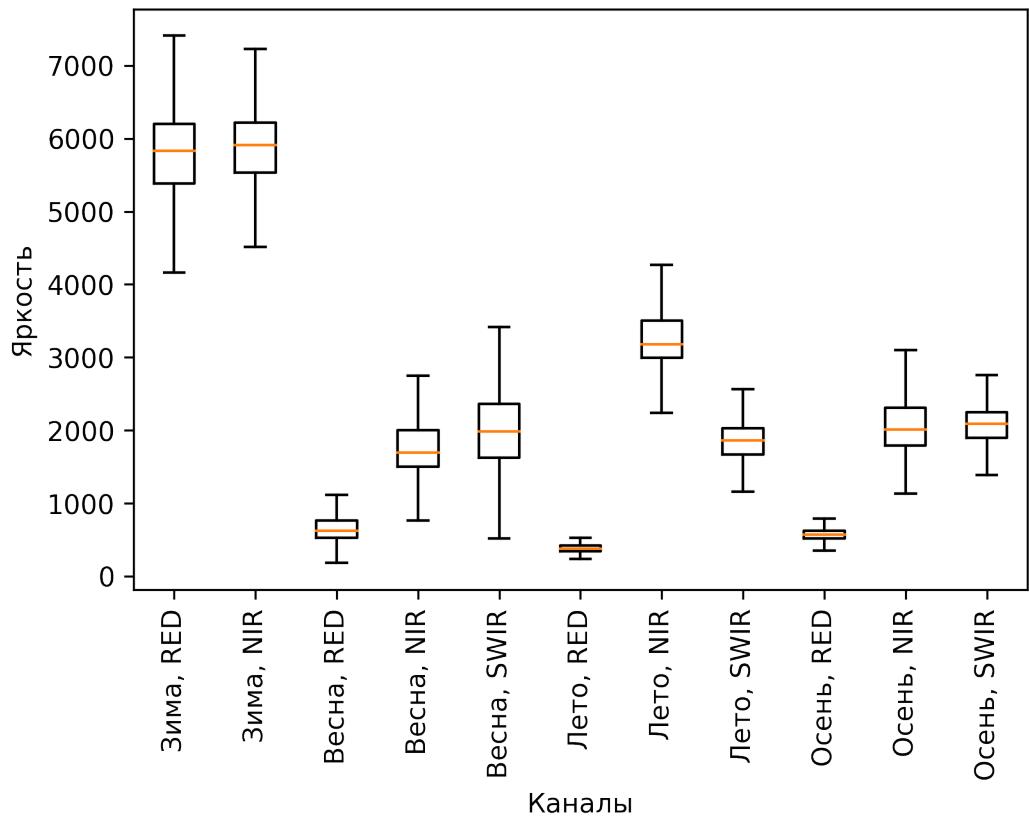
Травянистая тундра



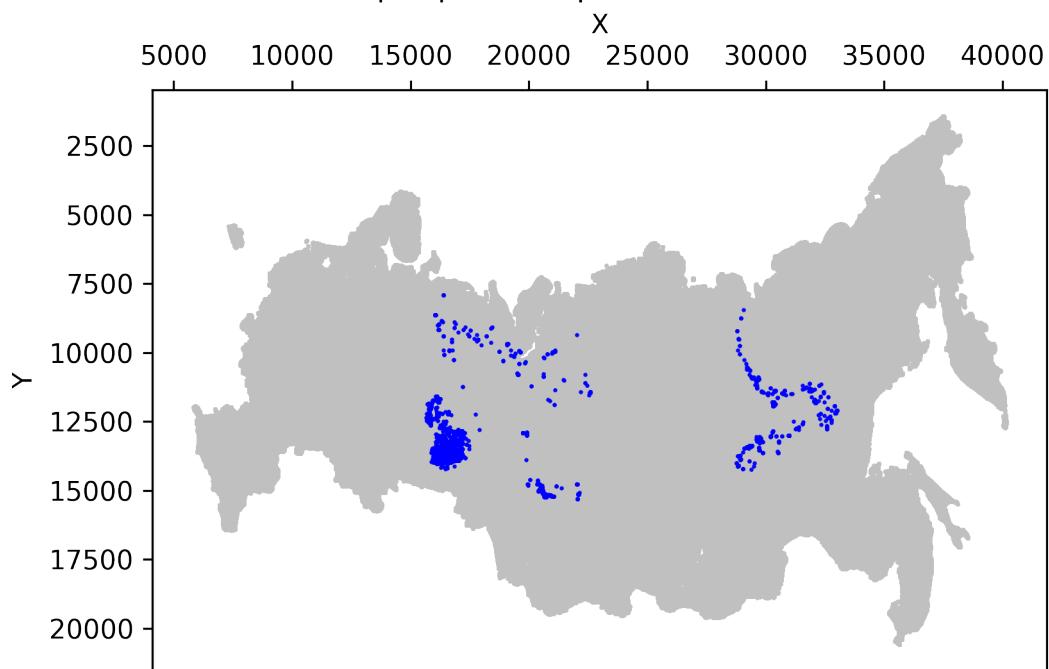




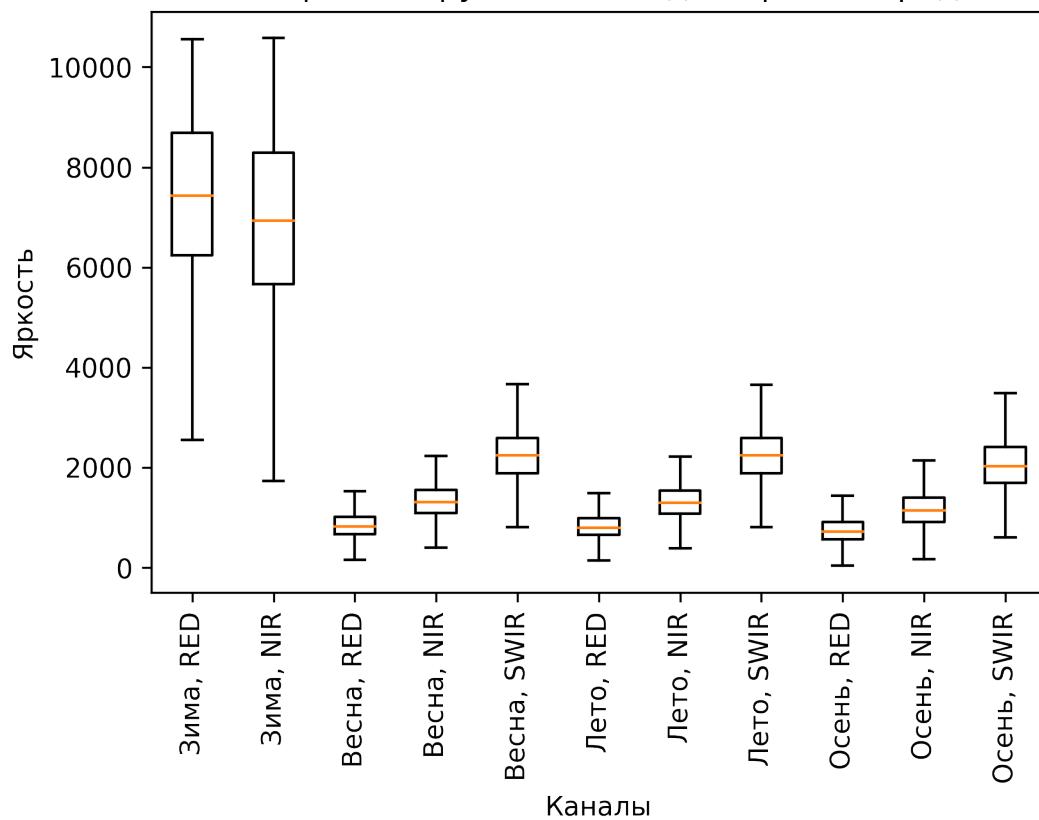
Прибрежная растительность



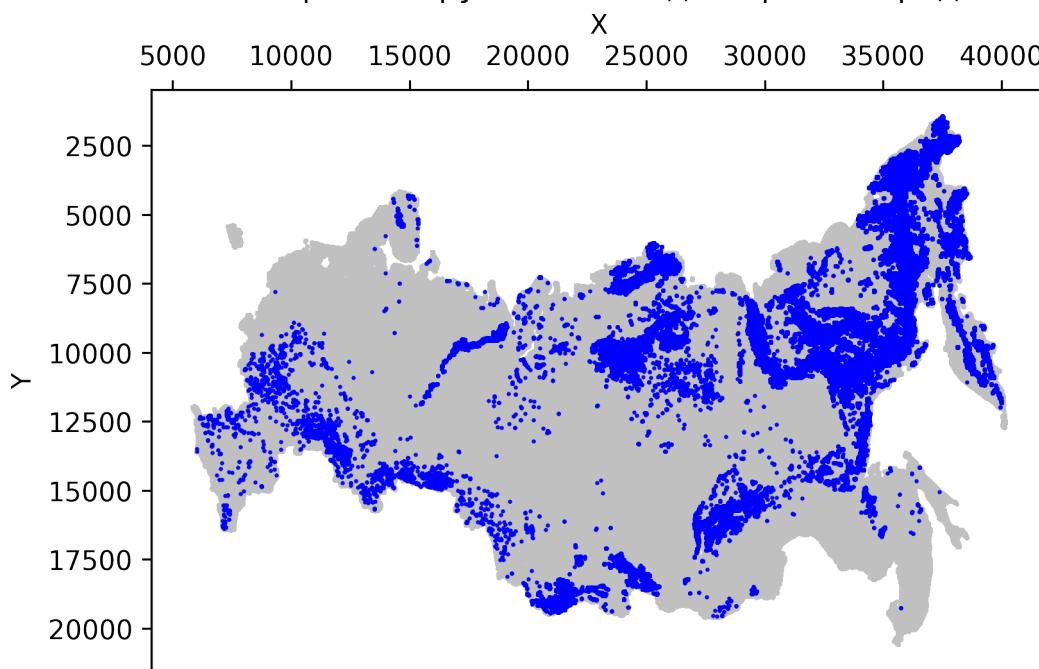
Прибрежная растительность



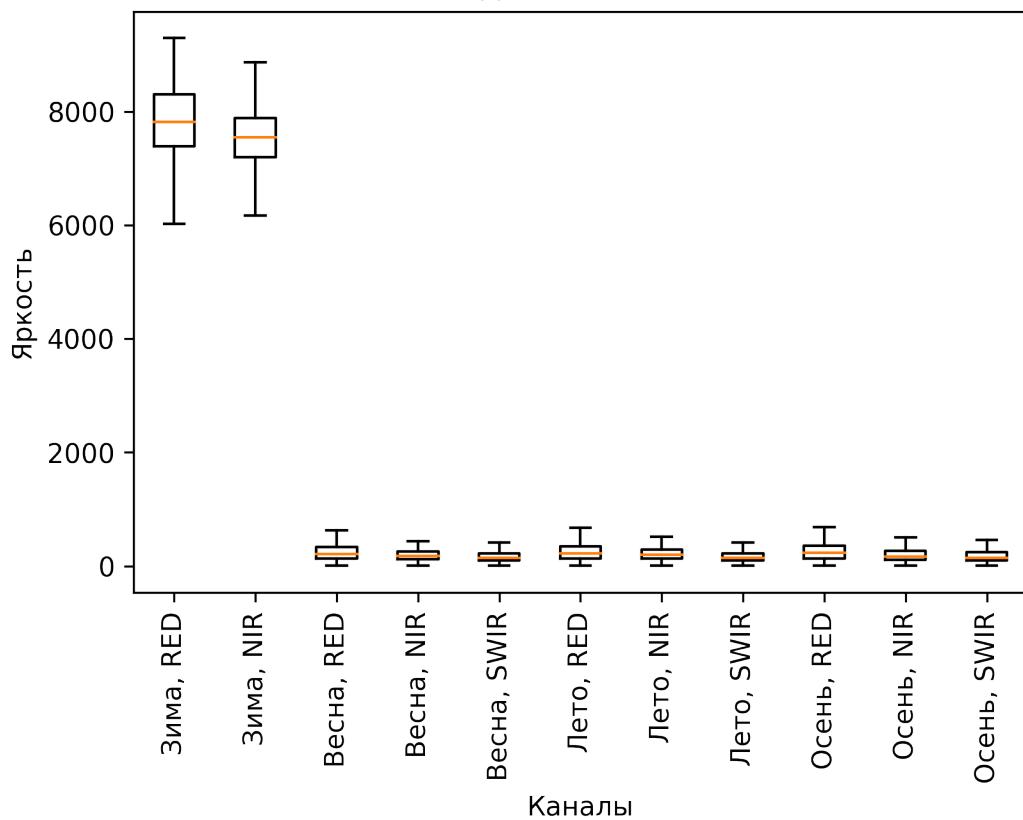
Открытые грунты и выходы горных пород



Открытые грунты и выходы горных пород



Водные объекты



Водные объекты

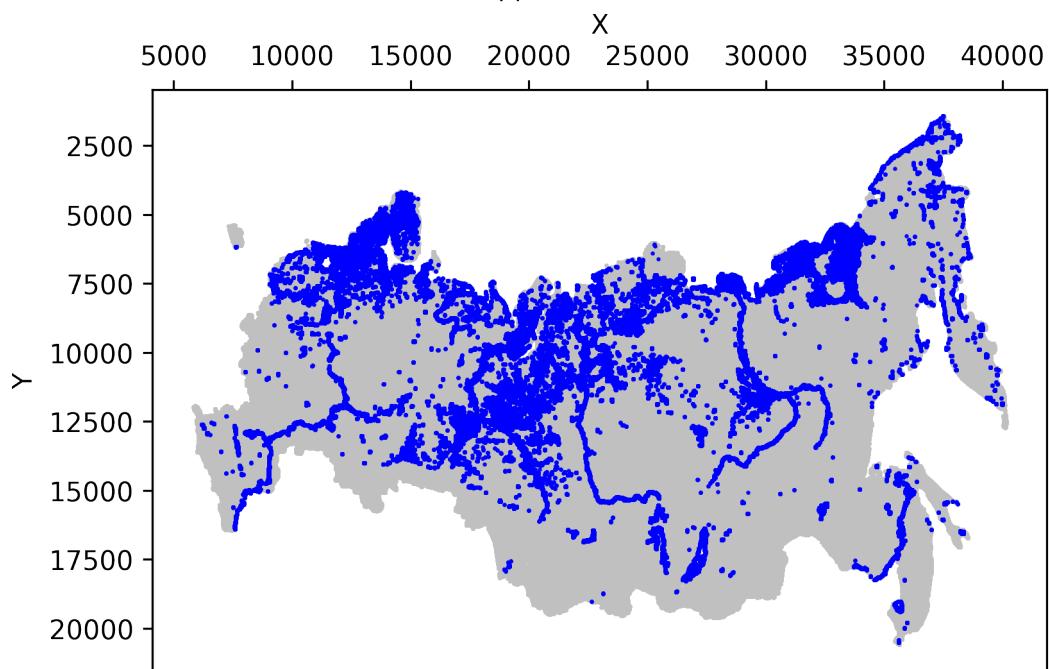


Рис. 3: Сравнение медианных значений сезонных яркостей в различных спектральных каналах по тематическим классам земного покрова, представленным в выборке

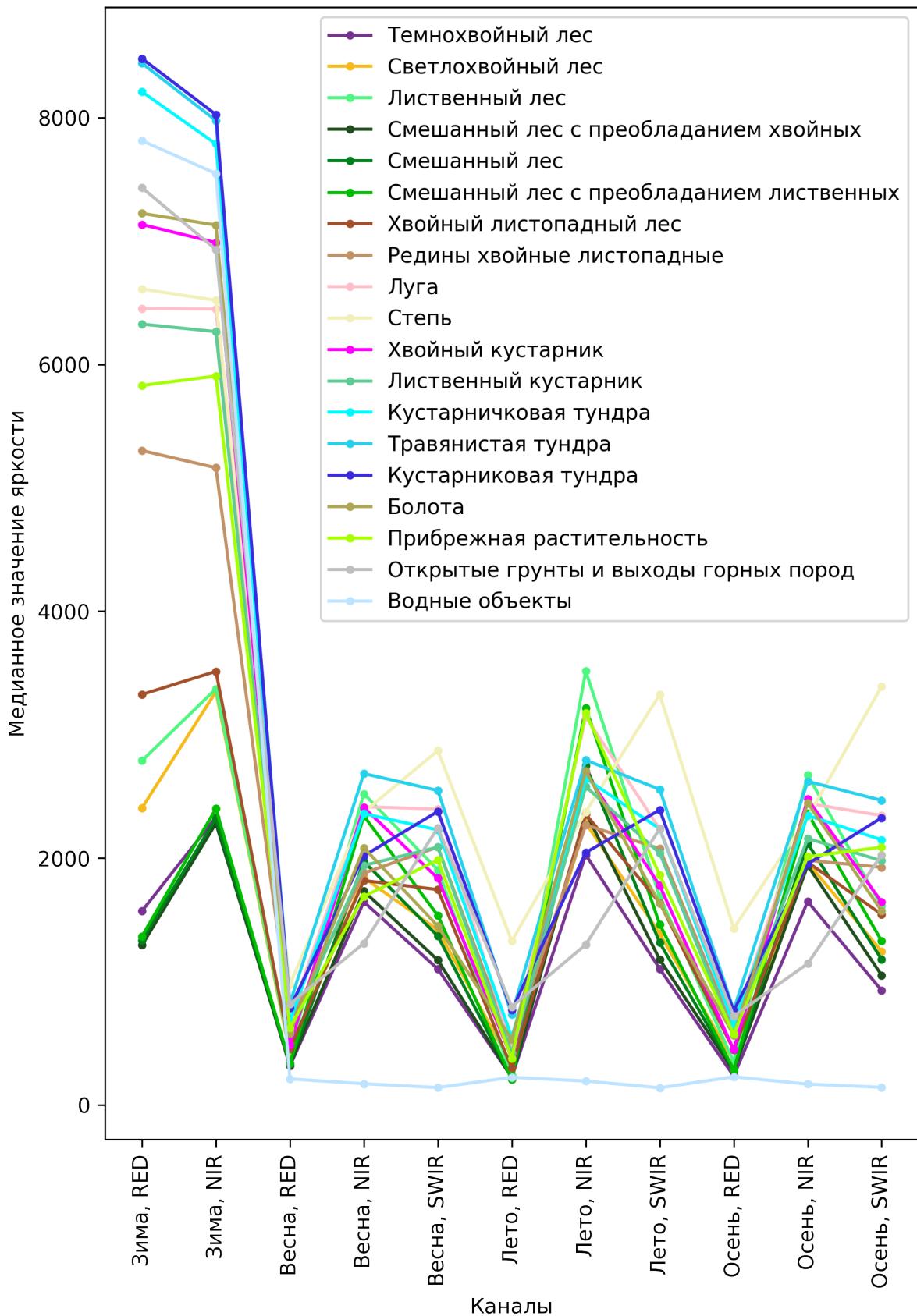
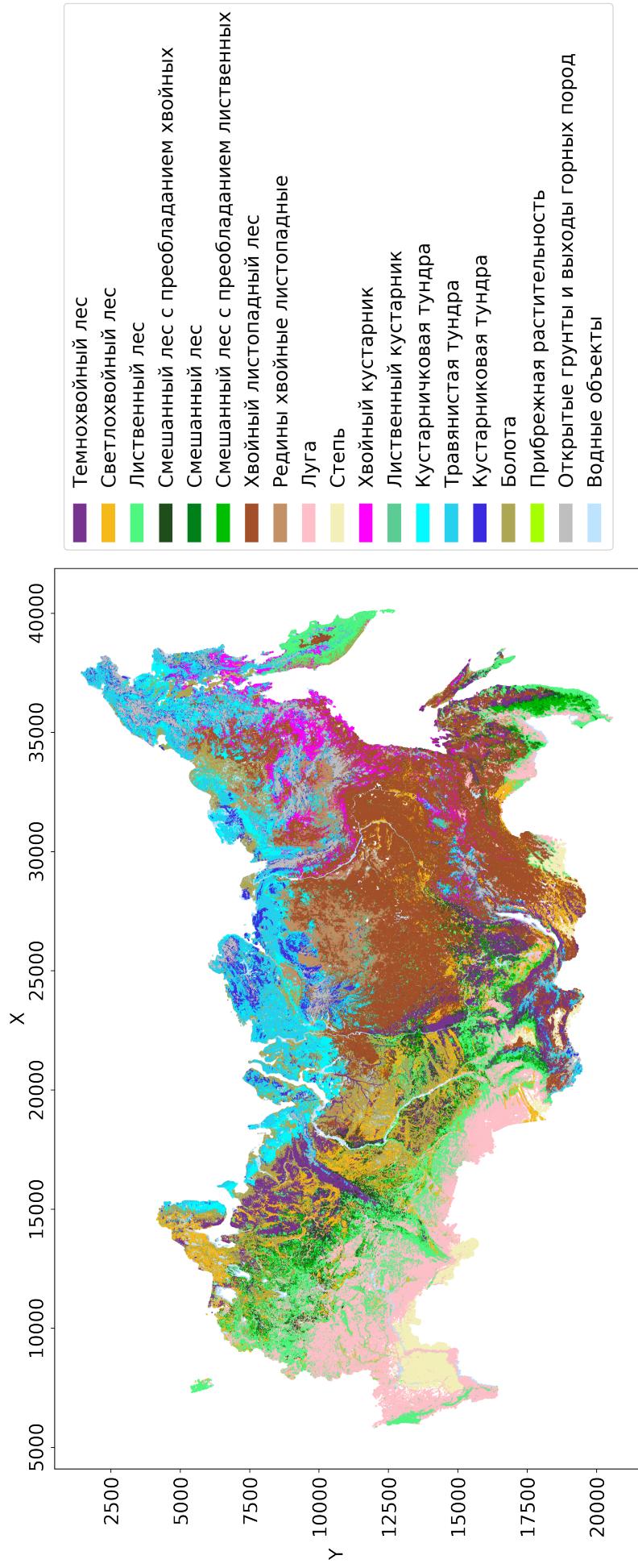


Рис. 4: Карта земного покрова России по данным, предоставленным в выборке



2.4 Инструменты для анализа данных

технические штуки: python, ПРОЧЕСТЬ ДОКУМЕНТАЦИЮ ваех,
набор данных больше размера ОЗУ, matplotlib

Ссылки на [4], [5], [6]

3 Классификация земного покрова с помощью модели случайного леса

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ задачи классификации, определение машинного обучения (с учителем)

Почему случайный лес?

предобработка – прув, что она не влияет на качество классификации,

Обучение, проверка на полном наборе данных (нет ли переобучения)

МЕТРИКИ ДЛЯ КЛАССИФИКАЦИИ, визуализировать ошибки на карте, CONFUSION MATRIX

ПОДБОР ОПТИМАЛЬНЫХ ГИПЕРПАРАМЕТРОВ

4 Значимость признаков

ЗНАЧИМОСТЬ ПРИЗНАКОВ

Методы оценки значимости признаков

Уменьшать количество признаков — оценивать влияние на качество классификации и время обучения

5 Применение модели случайного леса к полному набору данных

Применение результатов исследования к полному набору данных,
confusion matrix

6 Применение модели случайного леса к карте земного покрова

Получить табличные данные из изображений (GDAL)

Применить модель к полученным данным, применима ли модель?

7 Заключение

8 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 Vaex
<https://vaex.io/docs>
- [6] Документация библиотеки Matplotlib
<https://matplotlib.org/stable/>

9 Приложения

- ↗ Репозиторий проекта на GitHub

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