Классификация сельскохозяйственных культур, произрастающих на территории Хабаровского края на основе временных рядов значений оптических каналов и вегетационного индекса NDVI

ScienceDataLab

bylion

Шибаев Роман

Максименко Ксения

Структура данных

Features

Значение индекса NDVI (от -1 до +1) в день от начала года(название столбца - номер дня)

Target

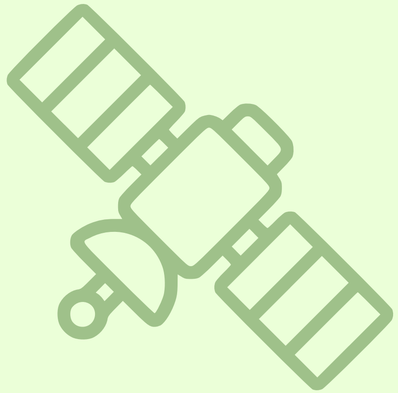
Culture - название выращиваемой культуры

1	залежь
2	соя
3	многолетние травы
4	зерновые
5	кукуруза
6	овощи

27 столбцов и 5874 строк

Данные по отдельным спутниковым каналам

B02	Blue
B03	Green
B04	Red
B05	VRE
B06	VRE
B07	VRE
B8A	NIR
B11	SWIR
B12	SWIR



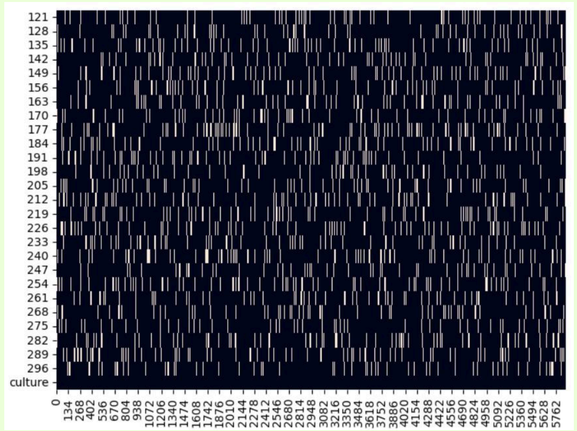
Предобработка данных

Обработка пропусков

121: 10.38%
128: 9.28%
135: 9.93%
142: 9.84%
149: 9.99%
156: 9.93%
163: 10.23%
170: 9.64%
177: 10.35%
184: 9.86%
191: 10.28%
198: 10.76%
205: 9.74%
212: 10.06%
219: 10.40%
226: 10.03%
233: 9.79%
240: 10.06%
247: 9.45%
254: 10.13%
261: 9.33%
268: 9.93%
275: 9.06%
282: 10.23%
289: 10.15%
296: 9.74%
culture: 0.00%

В среднем в каждой колонке 10% пропусков

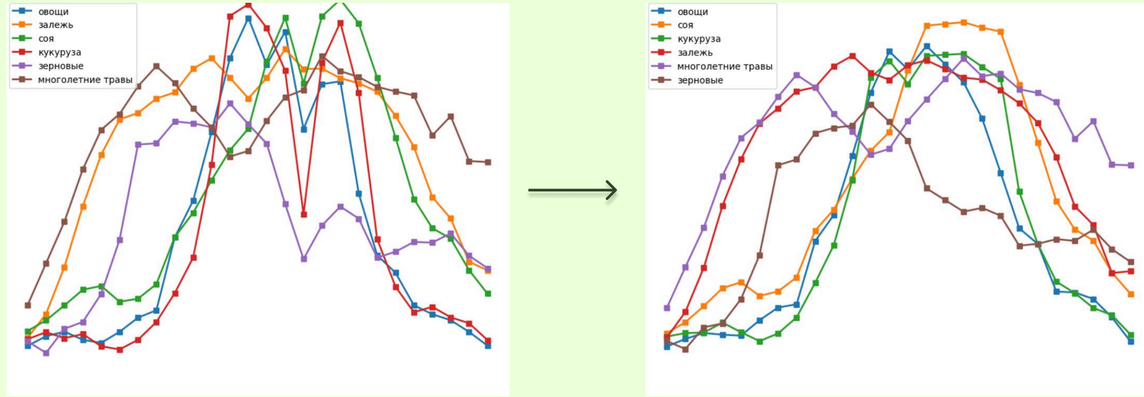
Структура данных - временной ряд, поэтому будем использовать **линейную интерполяцию**, как метод заполнения пропущенных значений



*Белое - пропущенные значения

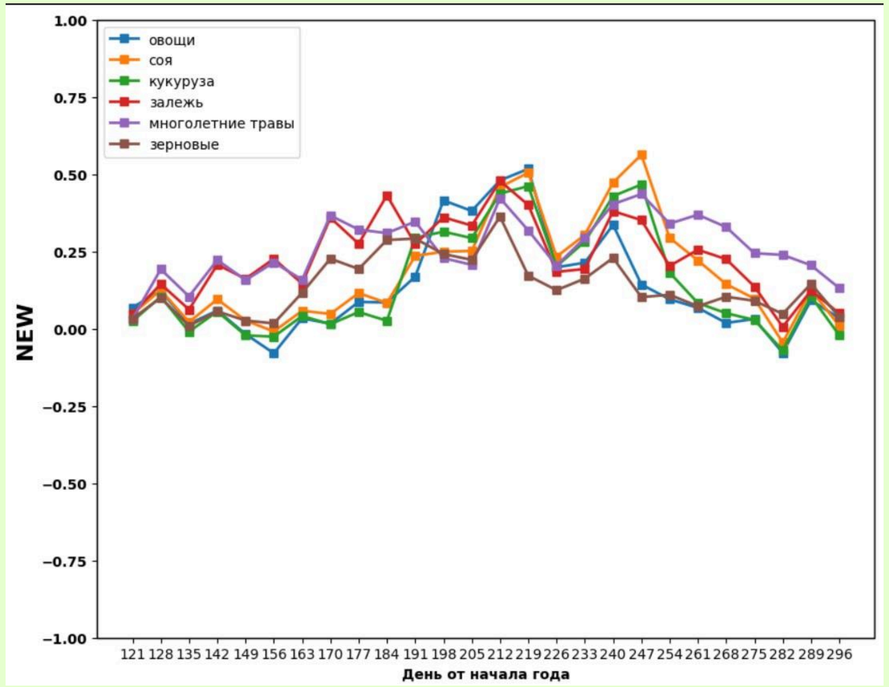
По визуализации можно заметить, что пропущенные значение не стоят группами подряд, а распределены по данным точечно, что ещё раз подтверждает, что мы можем использовать замену пропуска путем нахождения среднего соседних значений

Обработка выбросов

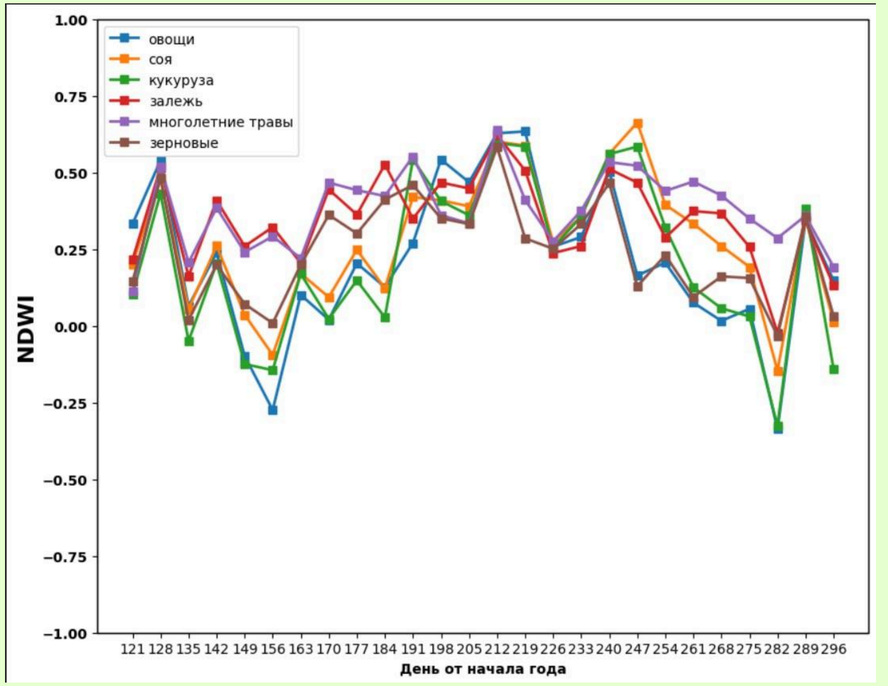


Дополнительные признаки

$$NEW = NDVI * NDWI$$



$$NDWI = \frac{NIR - SWIR}{NIR + SWIR}$$



Выбор модели

	Keras Model	Decision Tree	Random Forest
Interpolate	0.98	0.91	0.94
Dropna	0.89	0.74	0.75
Mean	0.96	0.85	0.89

Анализ данных

1 У некоторых видов сельскохозяйственных культур похожие распределения, что может ухудшить работу модели в дальнейшем

2 Присутствуют выбросы, например в 226 день от начала года

3 Никакой из классов не имеет распределения близкого к нормальному

Временной ряд культур по среднему INDVI