INAF Usecase

Han Tran
htran@know-center.at

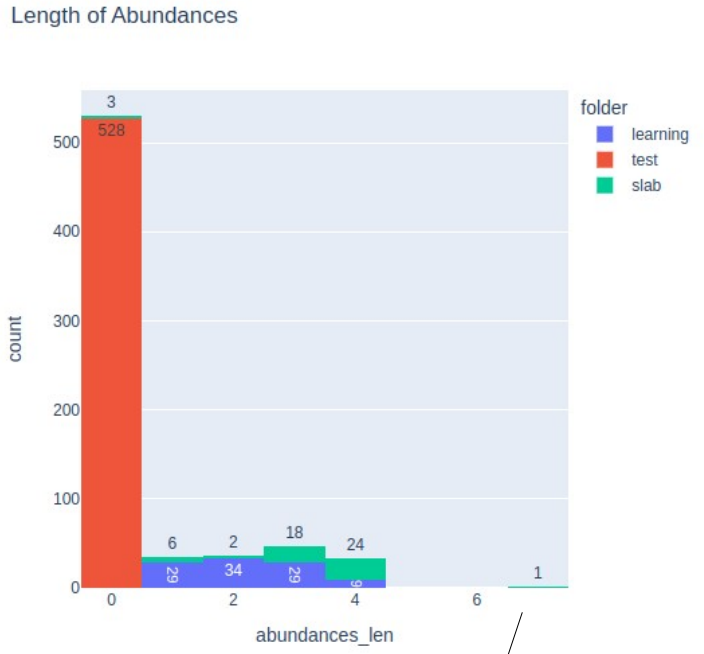
Data

	folder	file	wavelength	reflectance	error	abundances
0	learning	c1dl88a.json	[300.0, 301.0, 302.0, 303.0, 305.0, 306.0, 307...	[0.13368, 0.12945, 0.12522, 0.12099, 0.11253, ...	[0.09862, -1.0, -1.0, -1.0, 0.08009, -1.0, -1....	[{'mineral_phase_name': 'Clinopyroxene', 'perc...
1	learning	c1dl85a.json	[300.0, 301.0, 302.0, 303.0, 305.0, 306.0, 307...	[0.15799, 0.15419, 0.15039, 0.14659, 0.13898, ...	[0.09974, -1.0, -1.0, -1.0, 0.08056, -1.0, -1....	[{'mineral_phase_name': 'Clinopyroxene', 'perc...
2	learning	c1dd09.json	[300.0, 301.0, 302.0, 303.0, 305.0, 306.0, 307...	[0.0698, 0.07041, 0.07102, 0.07163, 0.07285, 0...	[0.0744, -1.0, -1.0, -1.0, 0.05007, -1.0, -1.0...	[{'mineral_phase_name': 'Olivine', 'percentage...
3	learning	c1kc11.json	[300.0, 301.0, 302.0, 303.0, 305.0, 306.0, 307...	[0.1154, 0.11552, 0.11565, 0.11577, 0.11603, 0...	[0.02815, -1.0, -1.0, -1.0, 0.0218, -1.0, -1.0...	[{'mineral_phase_name': 'Clinopyroxene', 'perc...
4	learning	c1dl53a.json	[300.0, 301.0, 302.0, 303.0, 305.0, 306.0, 307...	[0.05884, 0.05687, 0.0549, 0.05293, 0.04898, 0...	[0.19212, -1.0, -1.0, -1.0, 0.12735, -1.0, -1....	[{'mineral_phase_name': 'Clinopyroxene', 'perc...
...	...	...	...	...	...	...
678	slab	9pl31m2b.json	[350.0, 351.0, 352.0, 353.0, 354.0, 355.0, 356...	[0.277059281, 0.2677906159, 0.2614362657, 0.25...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	[{'mineral_phase_name': 'Clinopyroxene', 'perc...
679	slab	7pl23m2b.json	[350.0, 351.0, 352.0, 353.0, 354.0, 355.0, 356...	[0.2311249462, 0.2201541758, 0.2108131243, 0.2...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	[{'mineral_phase_name': 'Clinopyroxene', 'perc...
680	slab	5pl15m2b.json	[350.0, 351.0, 352.0, 353.0, 354.0, 355.0, 356...	[0.1201666836, 0.1304344926, 0.142866868, 0.14...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	[{'mineral_phase_name': 'Clinopyroxene', 'perc...
681	slab	2pl38m1b.json	[350.0, 351.0, 352.0, 353.0, 354.0, 355.0, 356...	[0.1076584303, 0.1136064223, 0.1123374596, 0.1...	[-1.0, -1.0, -1.0, -1.0, -1.0, -1.0, -1.0, -1....	[{'mineral_phase_name': 'Clinopyroxene', 'perc...
682	slab	3pl37m2a.json	[350.0, 351.0, 352.0, 353.0, 354.0, 355.0, 356...	[0.1984755743, 0.1548056661, 0.1642029231, 0.1...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	[{'mineral_phase_name': 'Clinopyroxene', 'perc...

683 rows × 6 columns

Total: 683
Label: 152
No label: 531

Labels – Statistics

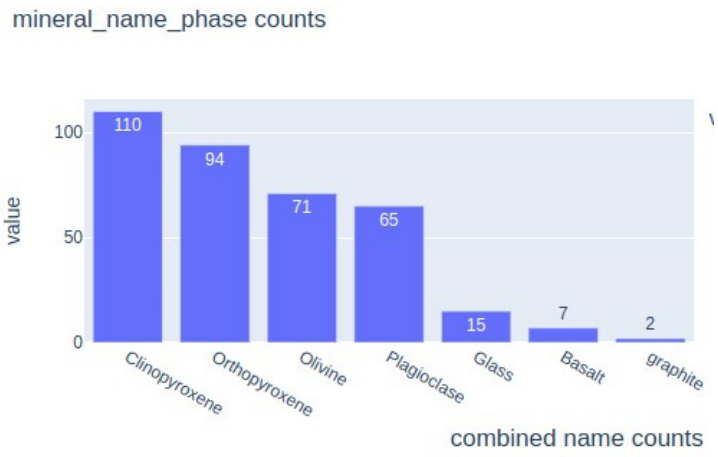


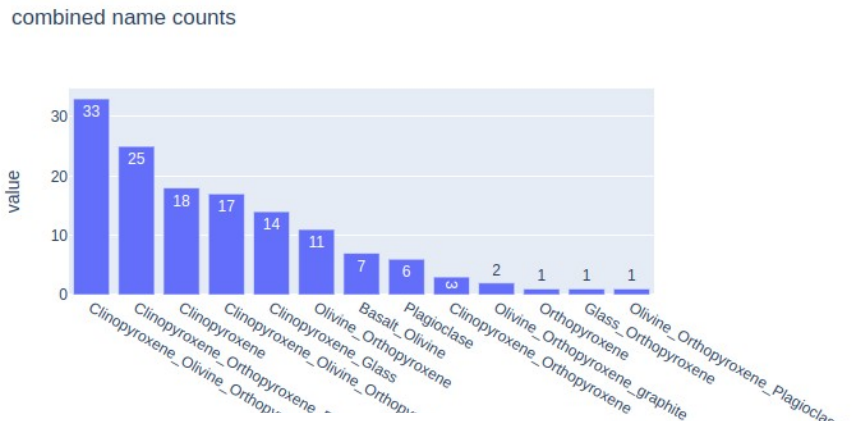
???

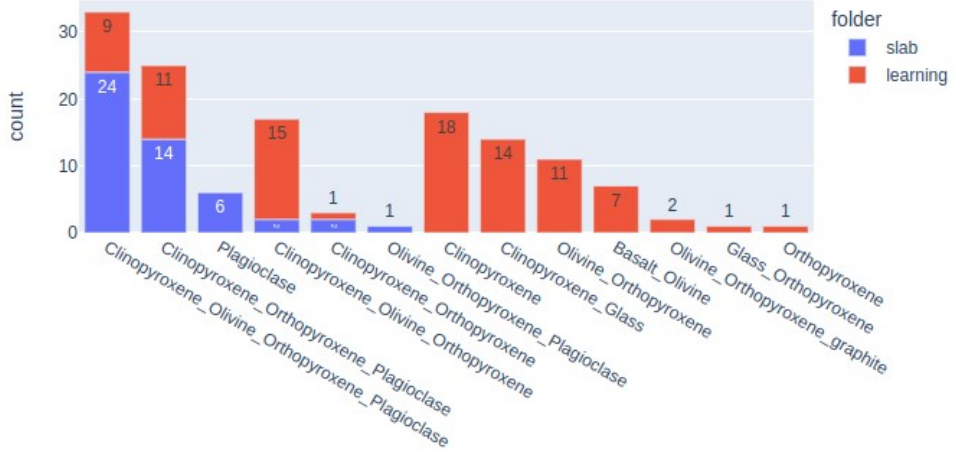
file	folder
5pl15m2a.json	slab

```
[[{'mineral_phase_name': 'Clinopyroxene', 'percentage': 2.0},
  {'mineral_phase_name': 'Olivine', 'percentage': 15.0},
  {'mineral_phase_name': 'Olivine', 'percentage': 34.0},
  {'mineral_phase_name': 'Plagioclase', 'percentage': 50.0},
  {'mineral_phase_name': 'Plagioclase', 'percentage': 50.0},
  {'mineral_phase_name': 'Orthopyroxene', 'percentage': 35.0},
  {'mineral_phase_name': 'Orthopyroxene', 'percentage': 14.0}]]
```

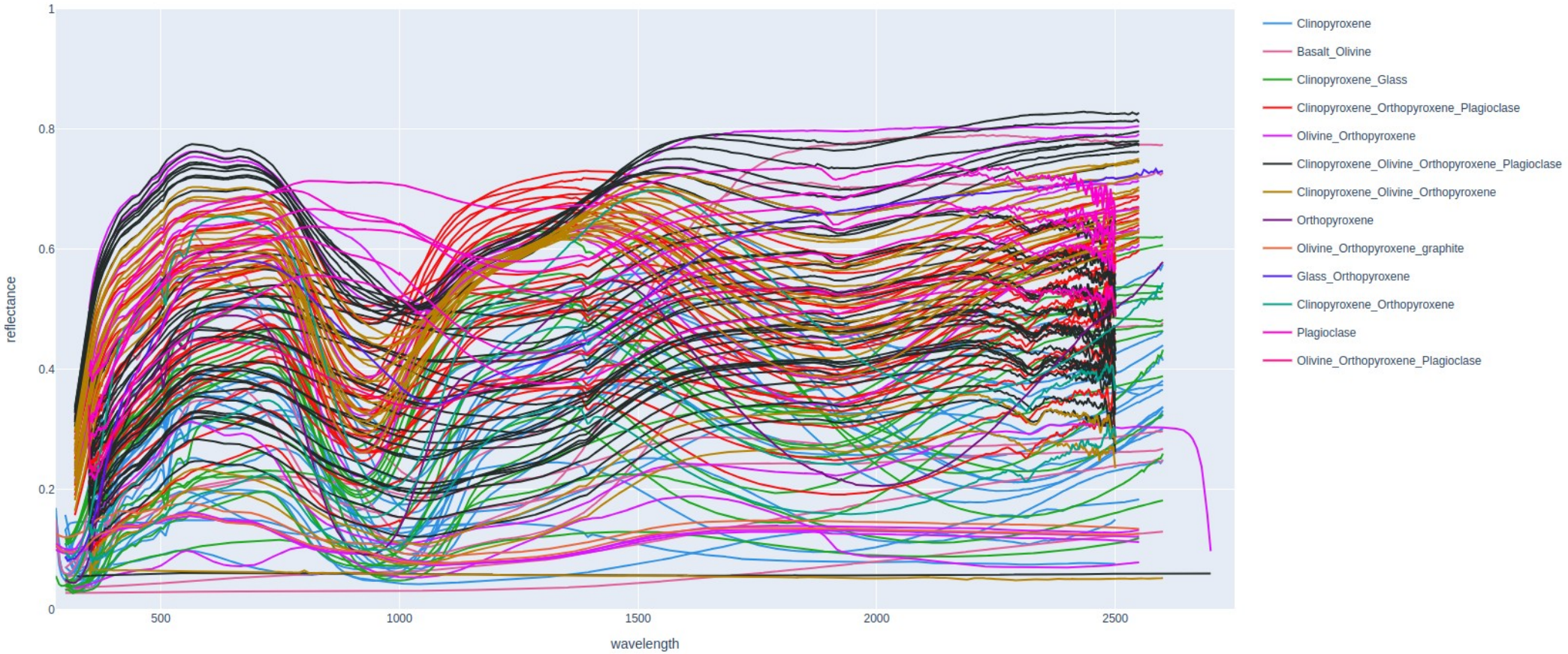
Labels – phase names (139)



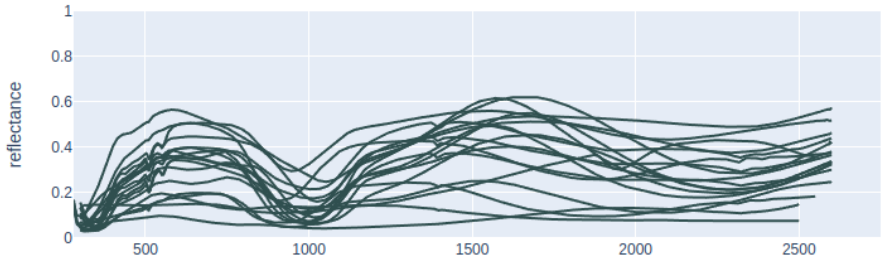




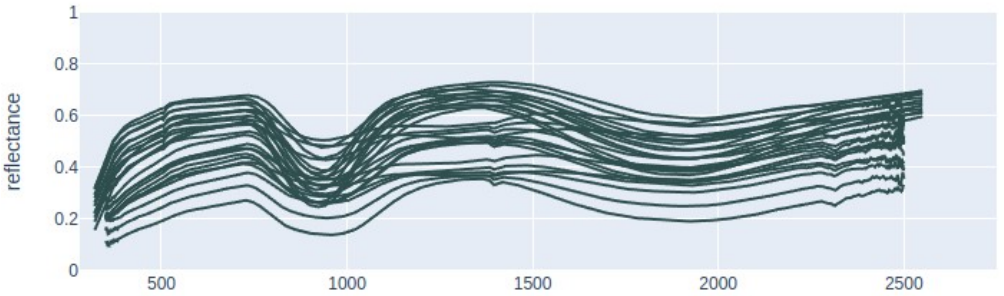
Labels – Combined Names



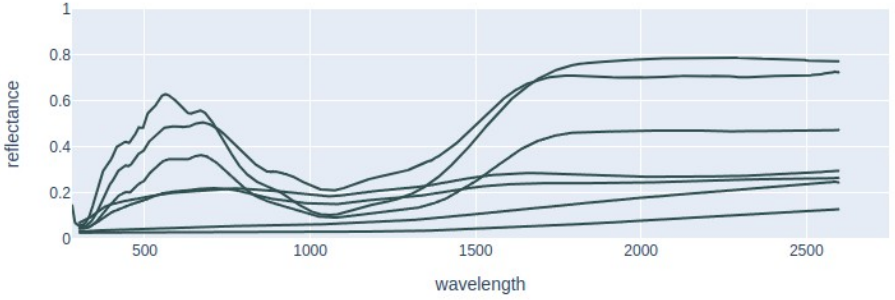
Clinopyroxene



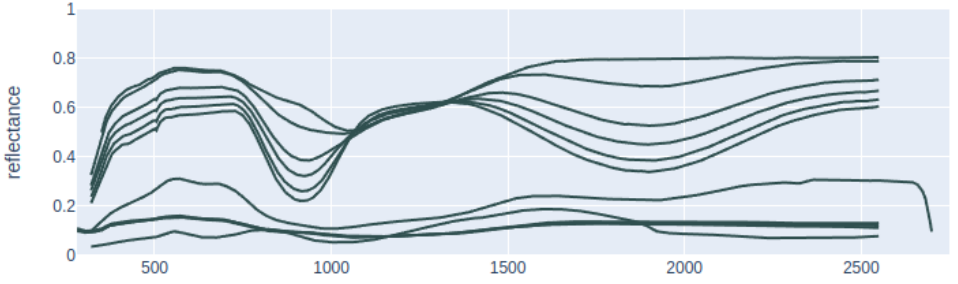
Clinopyroxene_Orthopyroxene_Plagioclase



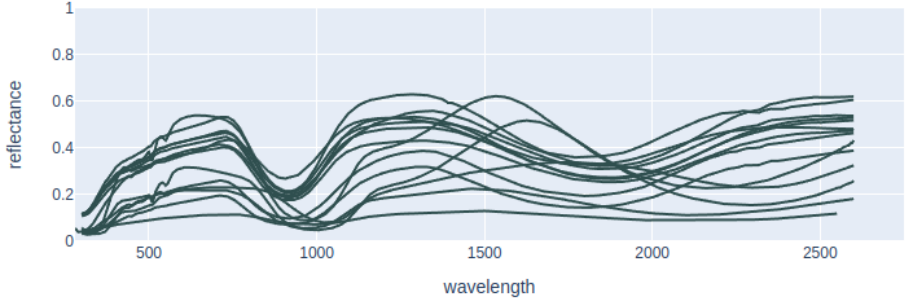
Basalt_Olivine



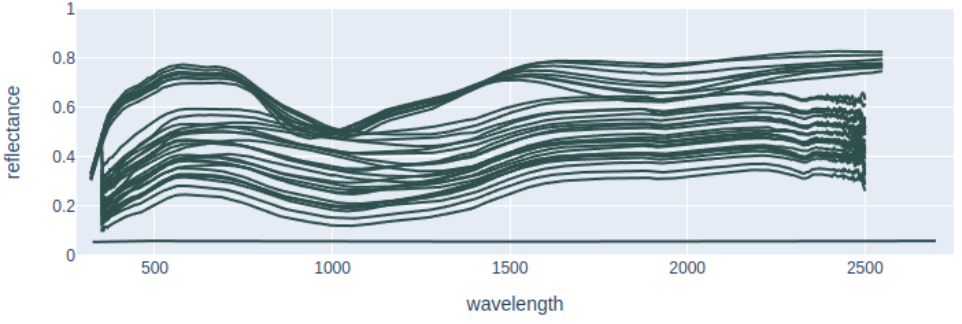
Olivine_Orthopyroxene



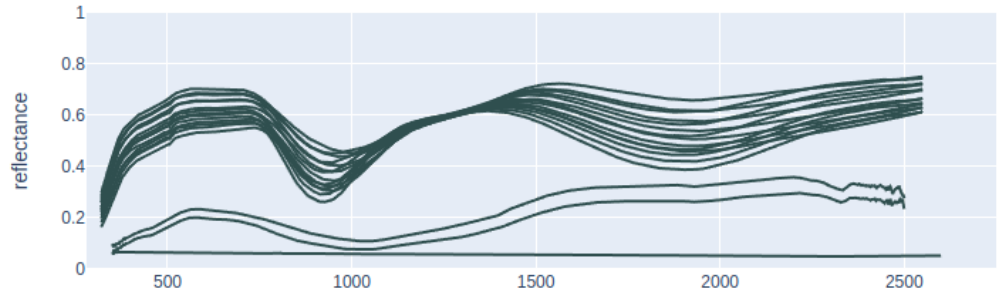
Clinopyroxene_Glass



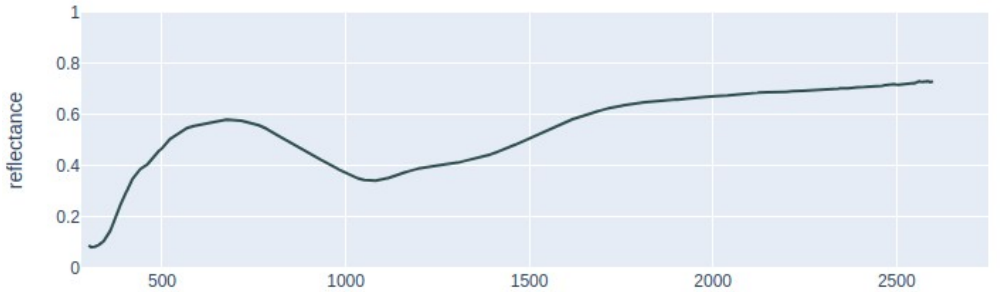
Clinopyroxene_Olivine_Orthopyroxene_Plagioclase



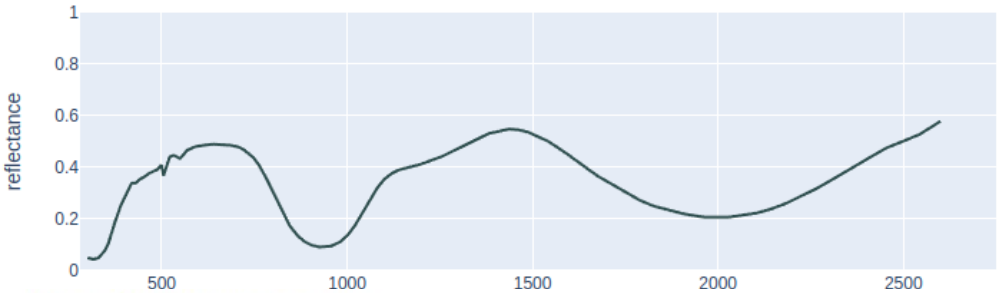
Clinopyroxene_Olivine_Orthopyroxene



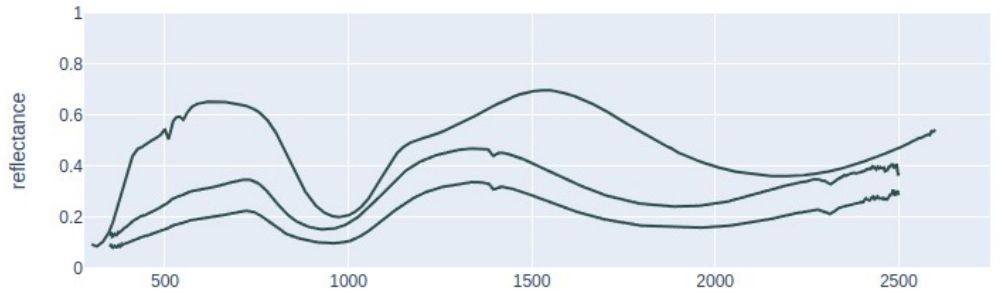
Glass_Orthopyroxene



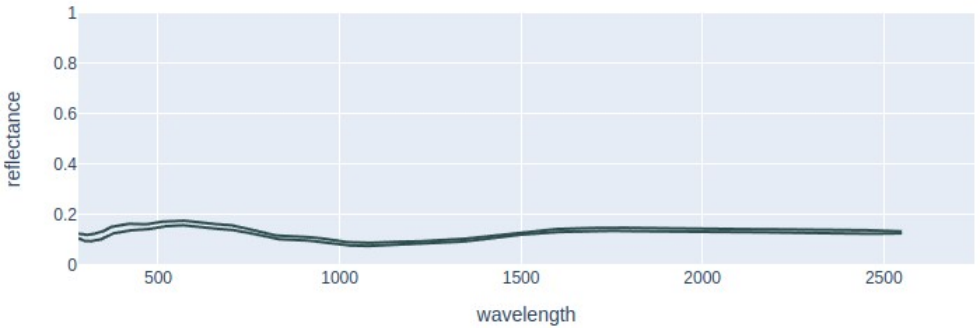
Orthopyroxene



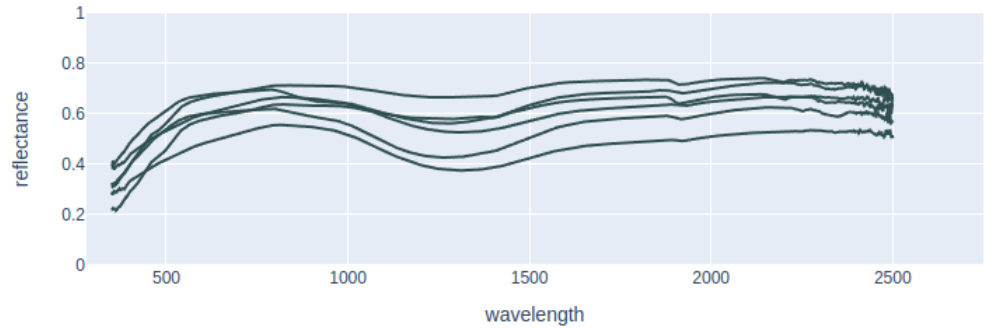
Clinopyroxene_Orthopyroxene

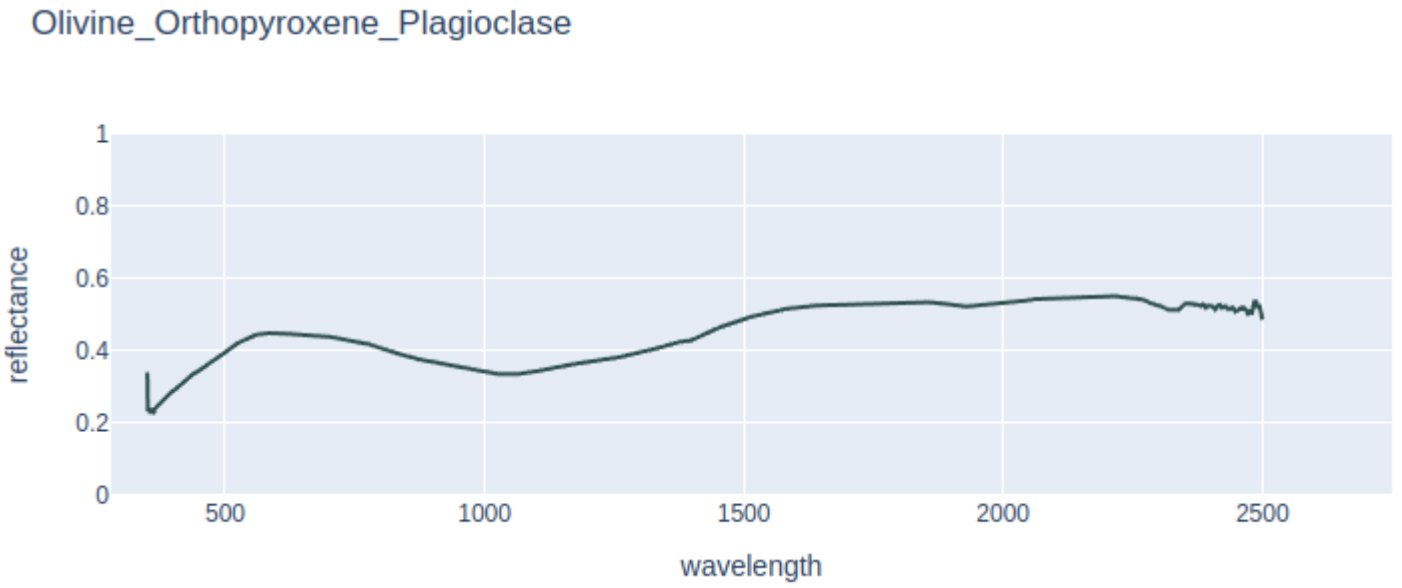


Olivine_Orthopyroxene_graphite



Plagioclase




Olivine_Orthopyroxene_Plagioclase
reflectance
1
0.8
0.6
0.4
0.2
0
500
1000
1500
2000
2500
wavelength

Duplicated ‘mineral_phase_name’

folder	file
learning	c1dl86a.json

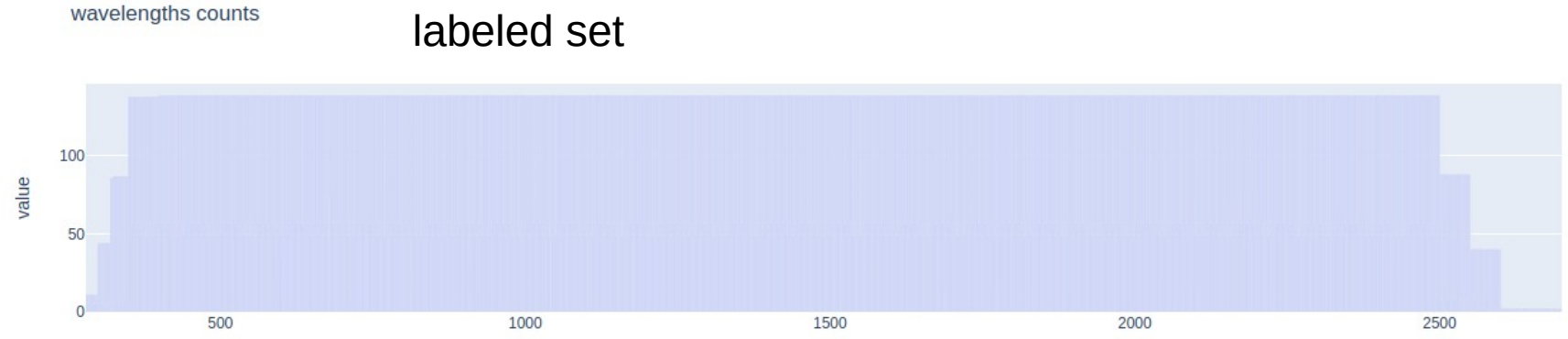
```
[[{'mineral_phase_name': 'Clinopyroxene', 'percentage': 50.0},
  {'mineral_phase_name': 'Clinopyroxene', 'percentage': 45.0},
  {'mineral_phase_name': 'Glass', 'percentage': 5.0}]]
```

Labels – not 100%

11 files

	folder	file	abundances
14	learning	c1jb482.json	[{'mineral_phase_name': 'Clinopyroxene', 'percentage': 99.0}]
18	learning	c1dl83a.json	[{'mineral_phase_name': 'Clinopyroxene', 'percentage': 99.0}]
31	learning	c1jb478.json	[{'mineral_phase_name': 'Clinopyroxene', 'percentage': 99.7}]
38	learning	c1dl91a.json	[{'mineral_phase_name': 'Clinopyroxene', 'percentage': 98.0}]
57	learning	c1jb483.json	[{'mineral_phase_name': 'Clinopyroxene', 'percentage': 97.0}]
60	learning	c1dl63a.json	[{'mineral_phase_name': 'Clinopyroxene', 'percentage': 99.5}]
76	learning	c1dl90a.json	[{'mineral_phase_name': 'Clinopyroxene', 'percentage': 70.0}]
91	learning	c1jb476.json	[{'mineral_phase_name': 'Clinopyroxene', 'percentage': 95.0}]
93	learning	c1jb485.json	[{'mineral_phase_name': 'Clinopyroxene', 'percentage': 95.0}]
99	learning	c1dl61a.json	[{'mineral_phase_name': 'Clinopyroxene', 'percentage': 99.5}]
649	slab	9pl21m1b.json	[{'mineral_phase_name': 'Clinopyroxene', 'percentage': 0.0}, {'mineral_phase_name': 'Plagioclase', 'percentage': 0.0}, {'mineral_phase_name': 'Orthopyroxene', 'percentage': 0.0}]

Spectrum – Wavelengths

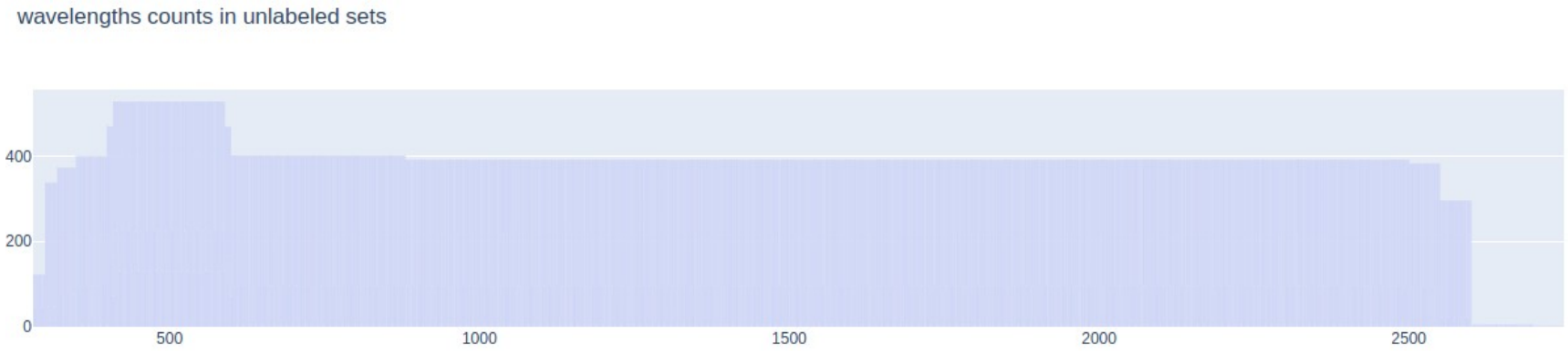


labeled set

```
e.index.min(), e.index.max()
```

```
(280, 2700)
```

```
folder
learning    90
slab        49
```



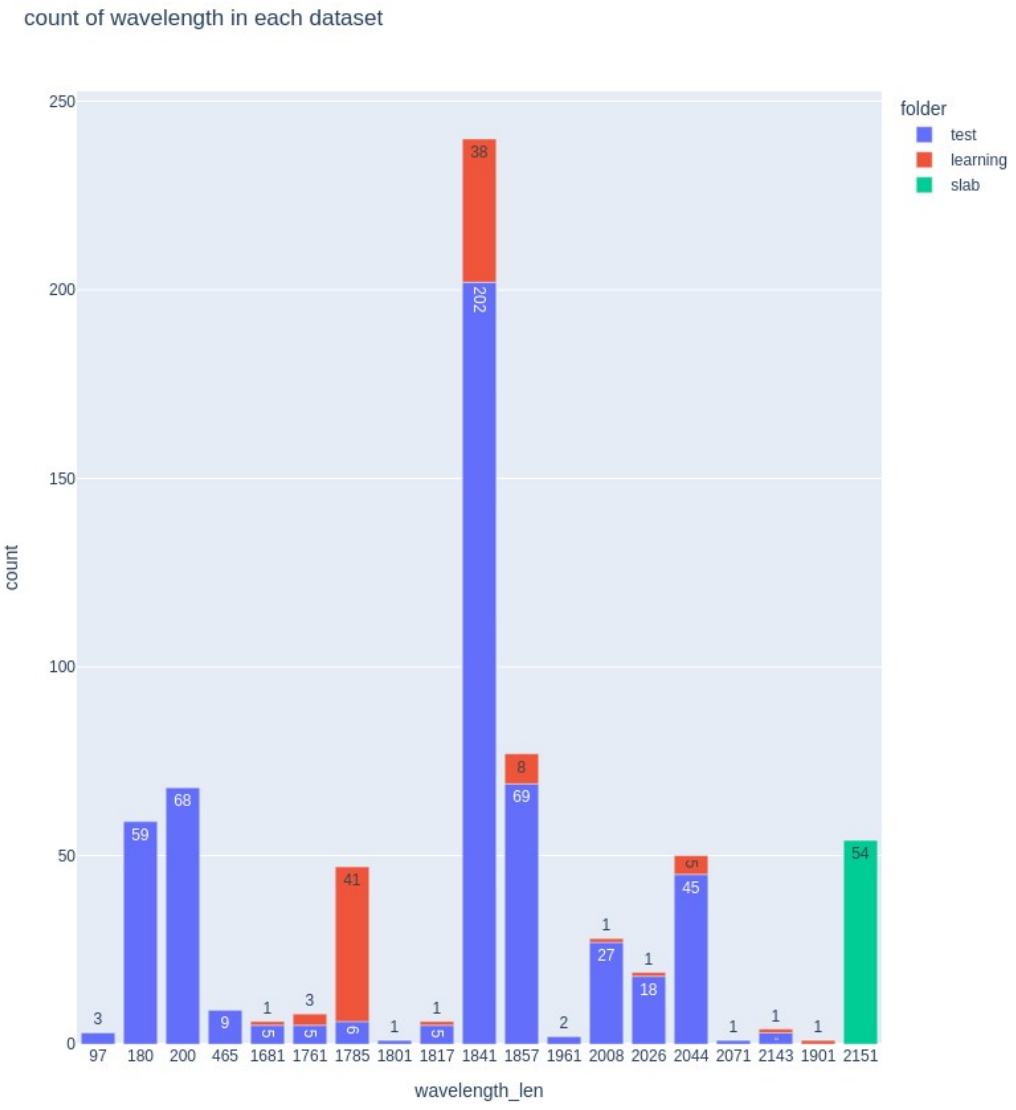

```
e.index.min(), e.index.max()
```

```
(280, 2750)
```

```
folder
test    528
slab      3
```

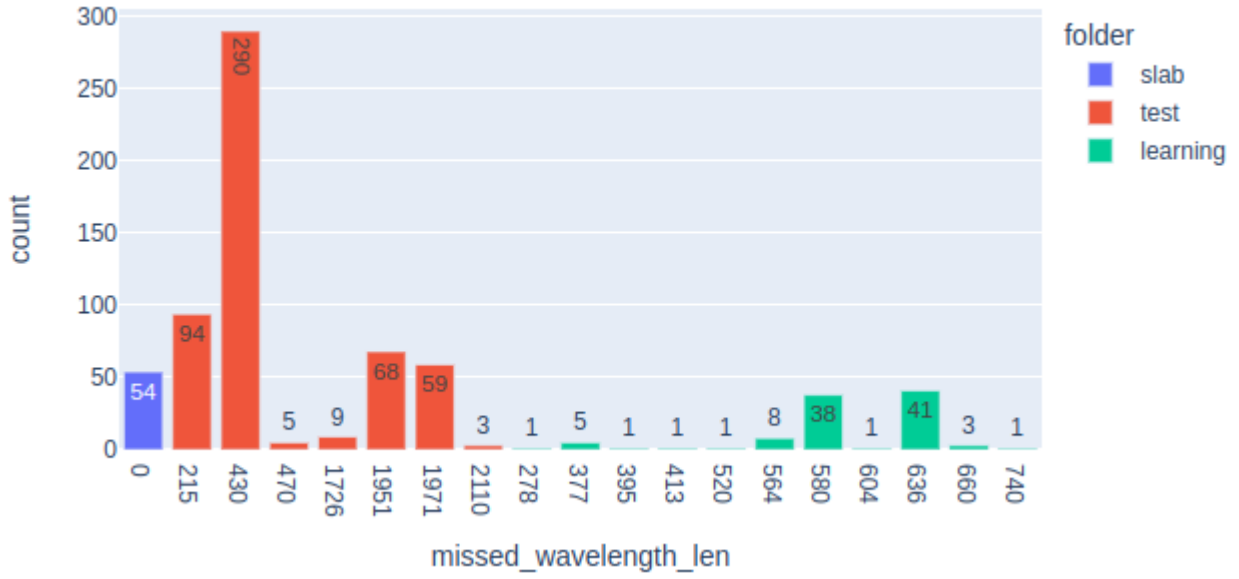
Length of wavelength in each sample *(or relfectance, or error)*

folder	wavelength_len	
learning	1681	1
	1761	3
	1785	41
	1817	1
	1841	38
	1857	8
	1901	1
	2008	1
	2026	1
	2044	5
	2143	1
slab	2151	54
test	97	3
	180	59
	200	68
	465	9
	1681	5
	1761	5
	1785	6
	1801	1
	1817	5
	1841	202
	1857	69
	1961	2
	2008	27
	2026	18
	2044	45
	2071	1
	2143	3

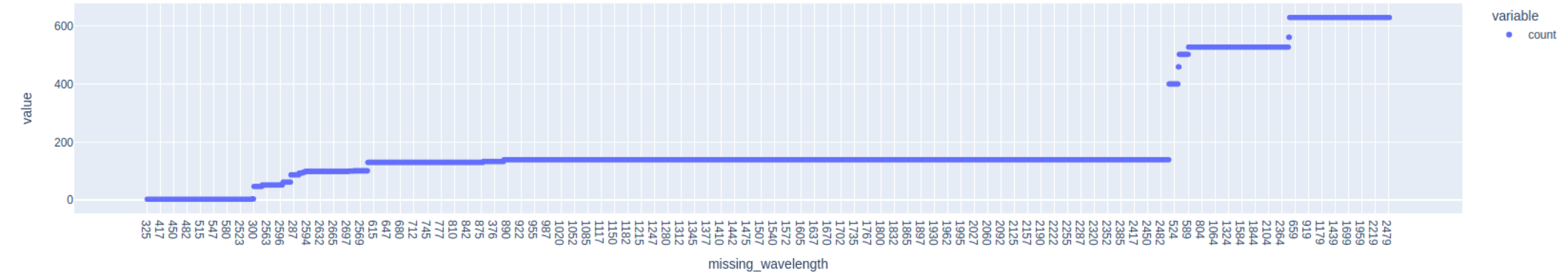


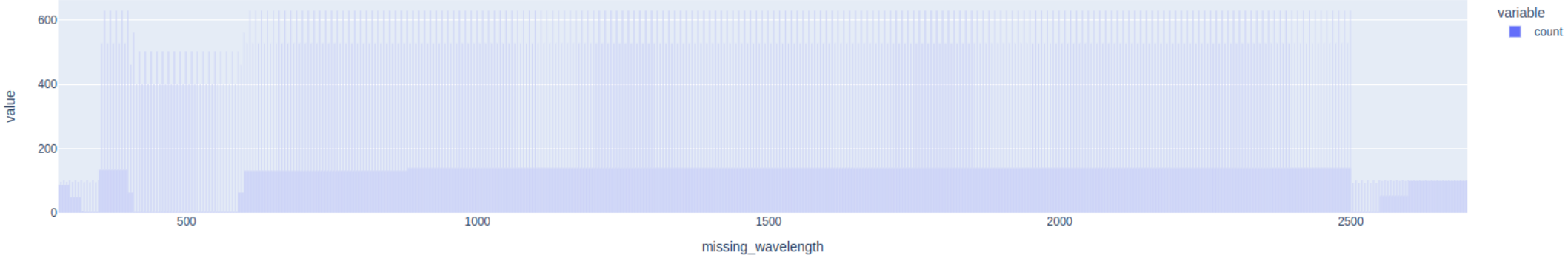
Missing Wavelengths

Missing Wavelength – the length of each sample



Missing Wavelength





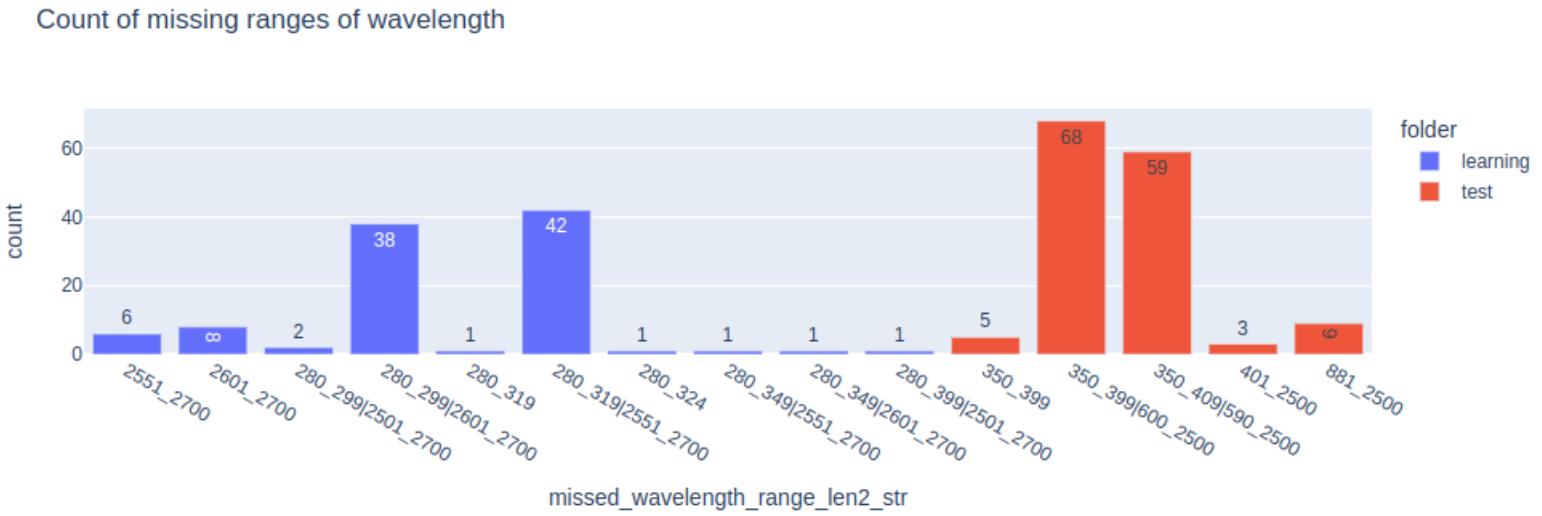
Missing Ranges

```
0 range(s) in missing wavelengths
missed_wavelength_len
430    290
215     94
0       54
Name: count, dtype: int64
========================
1 range(s) in missing wavelengths
missed_wavelength_len
1726    9
564     8
377     5
470     5
2110    3
278     1
520     1
604     1
Name: count, dtype: int64
========================
2 range(s) in missing wavelengths
missed_wavelength_len
1951    68
1971    59
636     41
580     38
660      3
740      1
413      1
395      1
Name: count, dtype: int64
```

290 samples have **430** single missing values

9 samples have **1** missing range and have total **1726** missing values

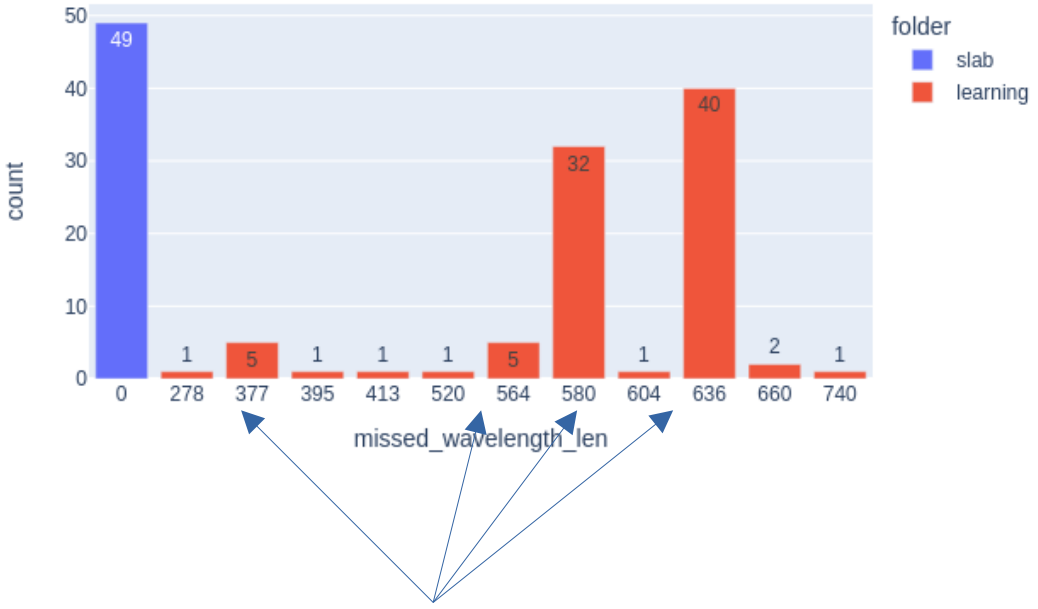
Missing Ranges



	index	count	#missing
0	401_2500	3	2100
1	590_2500	59	1911
2	600_2500	68	1901
3	881_2500	9	1620
4	2501_2700	3	200
5	2551_2700	49	150
6	280_399	1	120
7	2601_2700	47	100
8	280_349	2	70
9	350_409	59	60
10	350_399	73	50
11	280_324	1	45
12	280_319	43	40
13	280_299	40	20

Should take range
from 410 → 2500?

Missing Wavelength – train set



Same missing

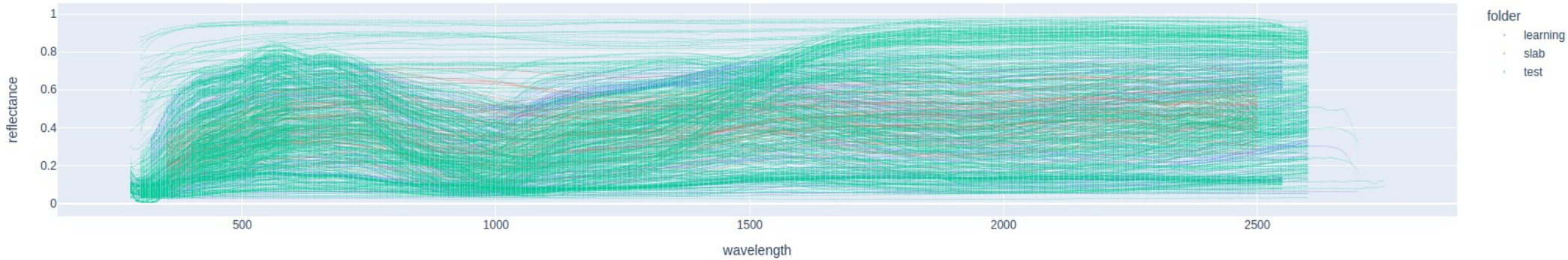
	folder	missed_wavelength_len	count
11	slab	0	49
5	learning	278	1
2	learning	377	5
6	learning	395	1
7	learning	413	1
8	learning	520	1
3	learning	564	5
1	learning	580	32
9	learning	604	1
0	learning	636	40
4	learning	660	2
10	learning	740	1

Reflectance

reflectance

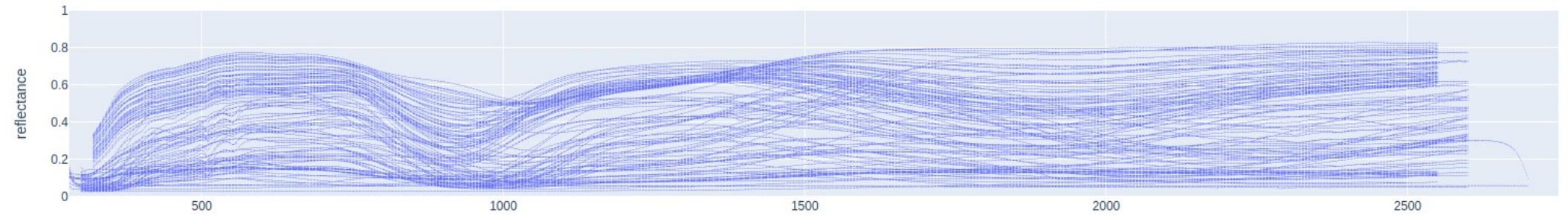
	min	max
learning	0.02712	0.8293
test	0.00328	0.98486

Range of wavelength vs. reflectance

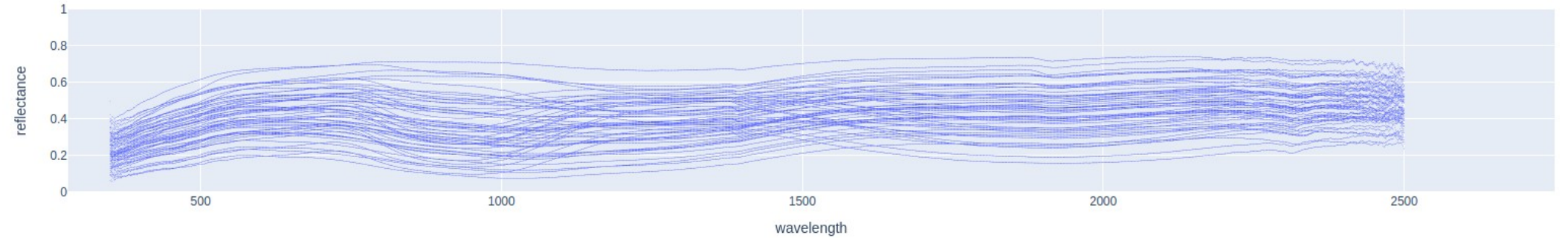


	wavelength		reflectance	
	min	max	min	max
folder				
learning	280.0	2700.0	0.027120	0.829300
slab	350.0	2500.0	0.058117	0.743495
test	280.0	2750.0	0.003280	0.984860

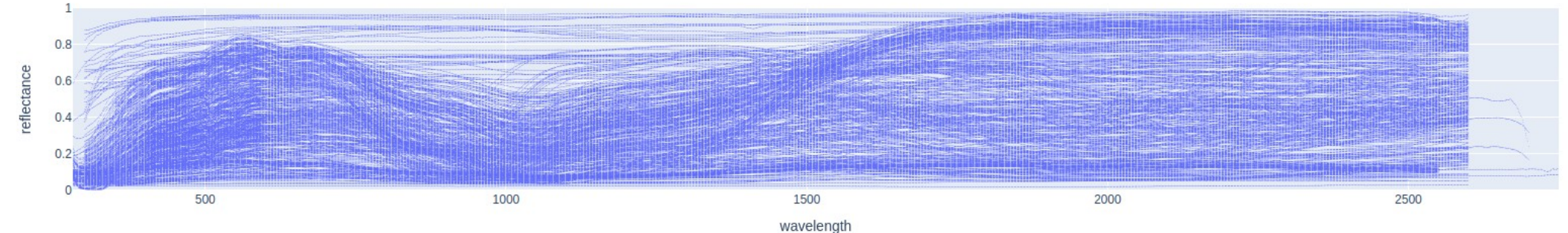
learning



slab



test



Other Questions

- error: The standard deviation of reflectance is calculated when applicable; otherwise, a value of -1.0 indicates that no error calculation was performed.

How about 0?

ML

Features (wavelength & reflectance)

Interpolate (rbf*)

Wavelength Range cut-off (410-2500)

Data

Dimension reduction (2151->20)

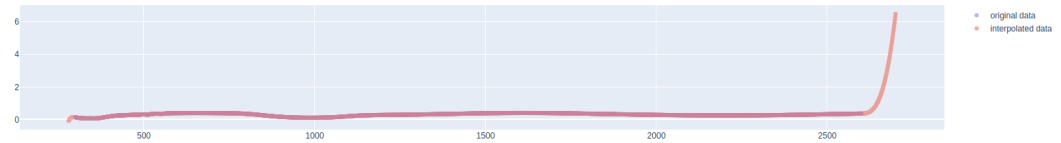
Label (abundances)

- Single Regressor
  - RandomForest
  - XGB
- MultioutputRegressor
  - RandomForest
  - XGB

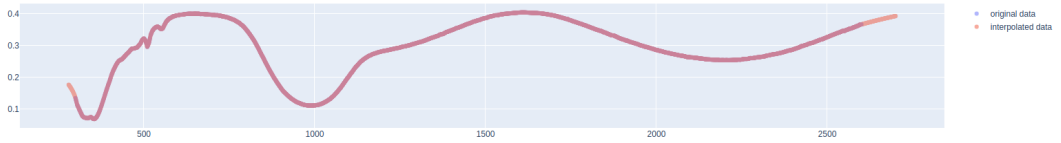
*https://docs.scipy.org/doc/scipy/reference/generated/scipy.interpolate.RBFInterpolator.html#scipy.interpolate.RBFInterpolator

Data Transformation – Features (wavelength & reflectance) - Interpolate (rbf)

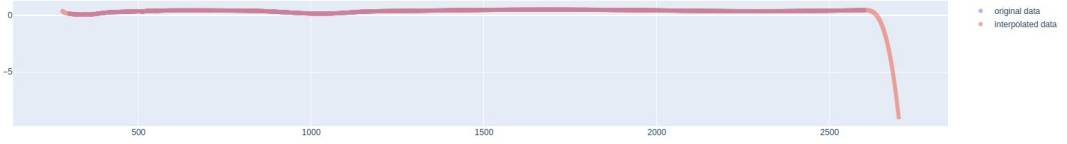
spline



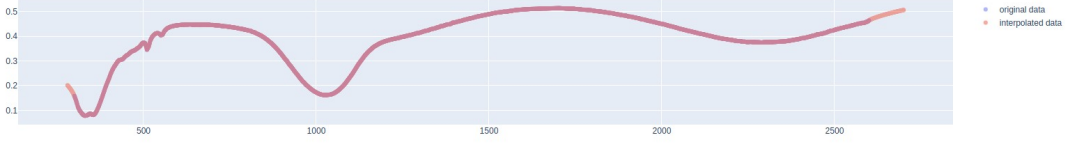
rbf



1
spline

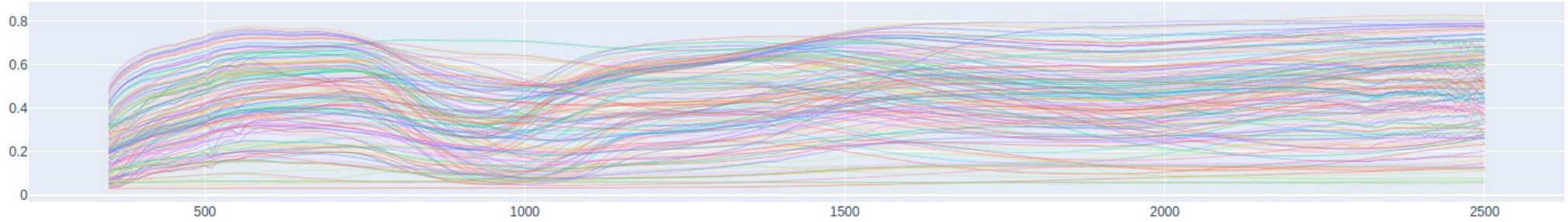


rbf



Check Appendix for more plots

Interpolated Reflectance



Results

Classification – RandomForest

Accuracy:
78.57 %

	Basalt_pred	Basalt_true	Clinopyroxene_pred	Clinopyroxene_true	Glass_pred	Glass_true	Olivine_pred	Olivine_true	Orthopyroxene_pred	Orthopyroxene_true	Plagioclase_pred	Plagioclase_true	graphite_pred	graphite_true
0	0	0	1	1	0	0	0	0	1	1	1	1	0	0
1	0	0	1	1	0	0	0	0	1	1	1	1	0	0
2	0	0	1	1	0	0	1	1	1	1	0	0	0	0
3	0	0	0	0	0	0	1	1	1	1	0	0	0	0
4	0	0	1	1	0	0	1	1	1	1	1	1	0	0
5	0	0	1	1	0	1	0	0	0	0	0	0	0	0
6	0	0	1	1	0	0	0	0	1	1	1	1	0	0
7	0	1	1	0	0	0	1	1	1	0	0	0	0	0
8	0	0	1	1	0	0	0	0	1	1	1	1	0	0
9	0	0	1	1	0	0	1	1	1	1	1	1	0	0
10	0	0	1	1	0	0	0	0	1	1	1	1	0	0
11	0	0	1	1	0	0	0	0	1	1	1	1	0	0
12	0	0	1	0	0	0	1	1	1	1	1	1	0	0
13	0	0	1	1	0	0	1	1	1	1	1	1	0	0
14	0	0	1	1	1	1	0	0	0	0	0	0	0	0
15	0	0	1	1	0	0	0	0	1	1	1	1	0	0
16	0	0	1	1	0	0	0	0	0	0	0	0	0	0
17	0	0	1	1	0	0	1	1	1	1	1	1	0	0
18	0	0	0	0	0	0	0	0	0	0	1	1	0	0
19	0	0	1	1	0	0	0	0	0	0	0	0	0	0
20	0	0	1	1	0	0	1	1	1	1	1	1	0	0
21	0	0	1	1	0	0	0	0	0	0	0	0	0	0
22	0	1	0	0	0	0	0	1	0	0	0	0	0	0
23	0	0	0	0	0	0	1	1	1	1	0	0	0	1
24	0	0	1	1	0	0	0	0	0	0	0	0	0	0
25	0	0	1	1	0	0	1	1	1	1	0	0	0	0
26	0	0	1	0	0	0	1	1	1	1	1	1	0	0
27	0	0	0	0	0	0	1	1	1	1	0	0	0	0

Classification – Multioutput RandomForest

Accuracy:
75 %

	Basalt_pred	Basalt_true	Clinopyroxene_pred	Clinopyroxene_true	Glass_pred	Glass_true	Olivine_pred	Olivine_true	Orthopyroxene_pred	Orthopyroxene_true	Plagioclase_pred	Plagioclase_true	graphite_pred	graphite_true
0	0	0	1	1	0	0	0	0	1	1	1	1	0	0
1	0	0	1	1	0	0	0	0	1	1	1	1	0	0
2	0	0	1	1	0	0	1	1	1	1	0	0	0	0
3	0	0	0	0	0	0	1	1	1	1	0	0	0	0
4	0	0	1	1	0	0	1	1	1	1	1	1	0	0
5	0	0	1	1	0	1	0	0	1	0	0	0	0	0
6	0	0	1	1	0	0	0	0	1	1	1	1	0	0
7	0	1	1	0	0	0	1	1	1	0	0	0	0	0
8	0	0	1	1	0	0	0	0	1	1	1	1	0	0
9	0	0	1	1	0	0	1	1	1	1	1	1	0	0
10	0	0	1	1	0	0	0	0	1	1	1	1	0	0
11	0	0	1	1	0	0	0	0	1	1	1	1	0	0
12	0	0	1	0	0	0	1	1	1	1	1	1	0	0
13	0	0	1	1	0	0	1	1	1	1	1	1	0	0
14	0	0	1	1	1	1	0	0	0	0	0	0	0	0
15	0	0	1	1	0	0	0	0	1	1	1	1	0	0
16	0	0	1	1	0	0	0	0	0	0	0	0	0	0
17	0	0	1	1	0	0	1	1	1	1	1	1	0	0
18	0	0	0	0	0	0	0	0	0	0	1	1	0	0
19	0	0	1	1	0	0	0	0	0	0	0	0	0	0
20	0	0	1	1	0	0	1	1	1	1	1	1	0	0
21	0	0	1	1	0	0	1	0	1	0	0	0	0	0
22	1	1	0	0	0	0	0	1	0	0	0	0	0	0
23	0	0	0	0	0	0	1	1	1	1	0	0	0	1
24	0	0	1	1	0	0	0	0	0	0	0	0	0	0
25	0	0	1	1	0	0	1	1	1	1	0	0	0	0
26	0	0	1	0	0	0	1	1	1	1	1	1	0	0
27	0	0	0	0	0	0	1	1	1	1	0	0	0	0

Regressor

	RMSE	MSE
y_multirf	11.537125	180.106920
y_rf	11.508910	187.641539
y_multixgb	10.274027	146.312465
y_xgb	10.816344	161.985059

Regression – RandomForest

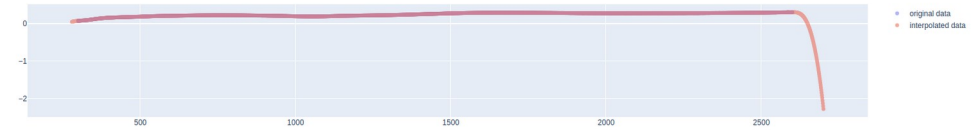
RMSE: 11.508910

	Basalt_pred	Basalt_true	Clinopyroxene_pred	Clinopyroxene_true	Glass_pred	Glass_true	Olivine_pred	Olivine_true	Orthopyroxene_pred	Orthopyroxene_true	Plagioclase_pred	Plagioclase_true	graphite_pred	graphite_true
0	1.50	0.0	9.7075	9.0	0.050	0.0	7.070	0.0	11.9225	11.0	69.75	80.0	0.00	0.0
1	0.75	0.0	38.3200	40.0	0.250	0.0	5.250	0.0	37.5300	40.0	17.90	20.0	0.00	0.0
2	0.00	0.0	7.5075	4.5	0.075	0.0	61.270	70.0	25.3975	25.5	5.75	0.0	0.00	0.0
3	1.85	0.0	8.8550	0.0	0.100	0.0	55.475	25.0	32.6700	75.0	0.20	0.0	0.85	0.0
4	0.10	0.0	11.0650	1.0	0.050	0.0	25.200	14.0	12.9350	5.0	50.65	80.0	0.00	0.0
5	1.90	0.0	54.8700	99.5	1.205	0.5	15.385	0.0	19.5900	0.0	6.90	0.0	0.15	0.0
6	0.40	0.0	17.0350	7.5	0.050	0.0	9.930	0.0	44.8850	42.5	27.70	50.0	0.00	0.0
7	5.60	95.0	16.3100	0.0	0.080	0.0	45.895	5.0	23.0150	0.0	8.65	0.0	0.45	0.0
8	0.10	0.0	25.3900	20.0	0.000	0.0	7.550	0.0	28.7600	20.0	38.20	60.0	0.00	0.0
9	0.30	0.0	5.0550	1.5	0.000	0.0	71.755	80.0	12.6900	8.5	10.15	10.0	0.05	0.0
10	0.00	0.0	23.5100	18.0	0.025	0.0	5.700	0.0	25.6650	22.0	45.10	60.0	0.00	0.0
11	0.00	0.0	16.2750	13.5	0.445	0.0	4.080	0.0	59.7000	76.5	19.50	10.0	0.00	0.0
12	0.00	0.0	1.3300	0.0	0.000	0.0	5.150	7.0	4.2200	3.0	89.30	90.0	0.00	0.0
13	2.70	0.0	4.1650	2.0	0.040	0.0	33.620	34.0	19.2750	14.0	40.20	50.0	0.00	0.0
14	1.35	0.0	74.0350	95.0	3.450	5.0	5.380	0.0	9.2850	0.0	6.50	0.0	0.00	0.0
15	0.00	0.0	14.7125	13.0	0.000	0.0	5.620	0.0	22.5175	17.0	57.15	70.0	0.00	0.0
16	0.70	0.0	77.0900	100.0	1.465	0.0	3.510	0.0	6.8350	0.0	10.40	0.0	0.00	0.0
17	0.20	0.0	5.5050	1.0	0.050	0.0	39.130	21.0	14.8650	8.0	40.25	70.0	0.00	0.0
18	0.05	0.0	6.4850	0.0	0.445	0.0	4.390	0.0	3.3300	0.0	85.30	100.0	0.00	0.0
19	0.00	0.0	91.6550	100.0	0.130	0.0	1.240	0.0	3.5250	0.0	3.45	0.0	0.00	0.0
20	0.00	0.0	4.2200	2.0	0.020	0.0	53.320	48.0	24.1400	20.0	18.30	30.0	0.00	0.0
21	1.50	0.0	58.4750	100.0	1.765	0.0	8.225	0.0	16.2350	0.0	13.75	0.0	0.05	0.0
22	16.90	90.0	21.4275	0.0	0.420	0.0	29.610	10.0	19.2425	0.0	12.40	0.0	0.00	0.0
23	1.00	0.0	11.8600	0.0	0.125	0.0	47.915	48.0	37.7500	48.0	0.20	0.0	1.15	4.0
24	1.90	0.0	68.3950	100.0	0.550	0.0	14.320	0.0	9.6850	0.0	5.15	0.0	0.00	0.0
25	0.00	0.0	33.2500	40.0	0.205	0.0	25.440	20.0	35.4050	40.0	5.70	0.0	0.00	0.0
26	0.10	0.0	9.7400	0.0	0.125	0.0	40.840	9.0	18.1950	21.0	31.00	70.0	0.00	0.0
27	0.50	0.0	11.2100	0.0	0.510	0.0	35.290	50.0	50.0900	50.0	2.40	0.0	0.00	0.0

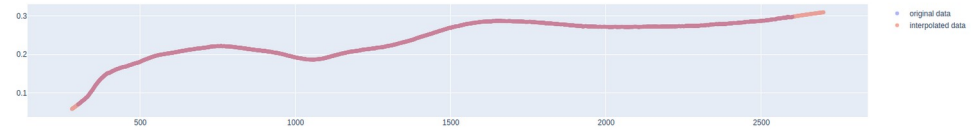
Thank you

Appendix – Interpolation Comparison

2
spline

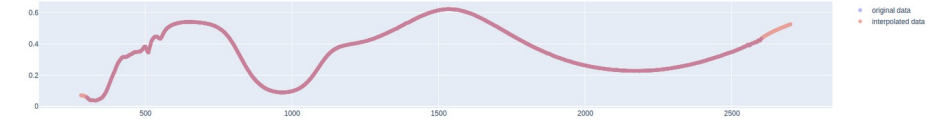


rbf

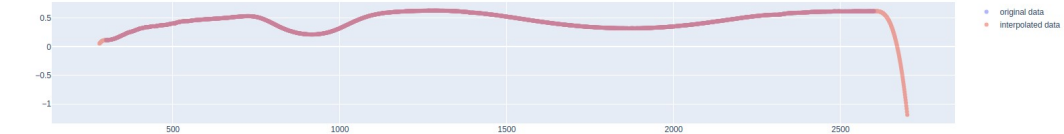


4
spline

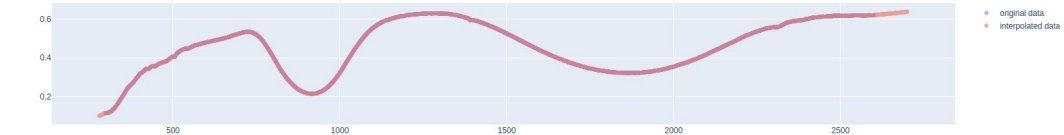
rbf



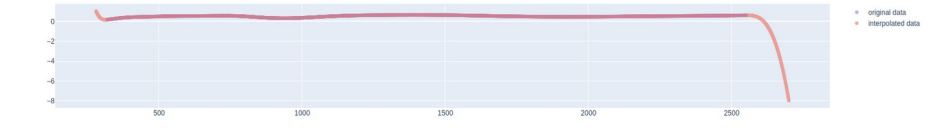
3
spline



rbf

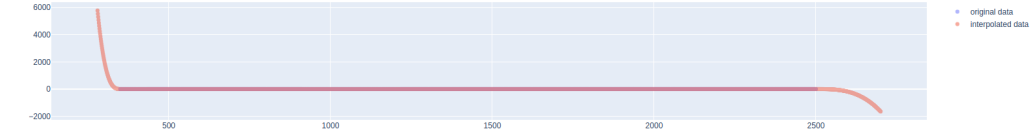


5
spline

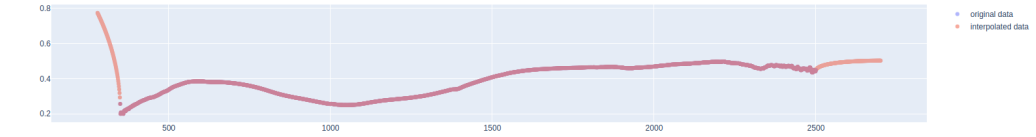


rbf

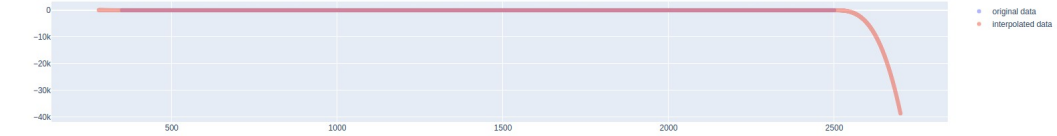
130
spline



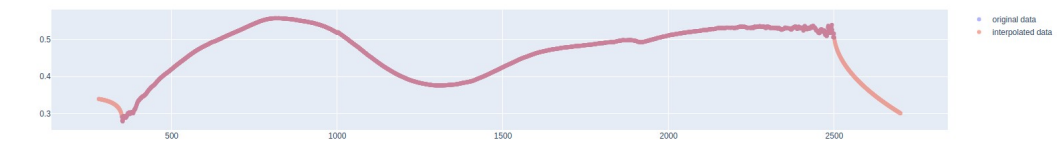
rbf



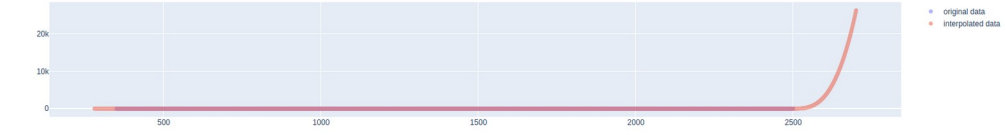
131
spline



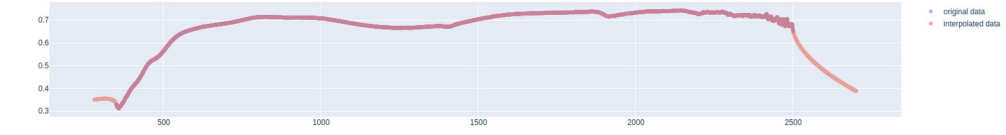
rbf



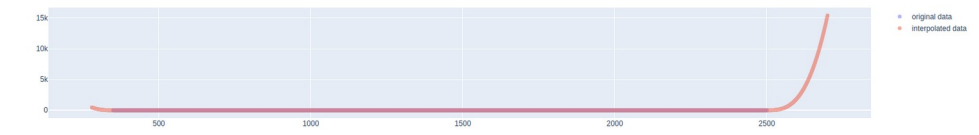
132
spline



rbf



133
spline



rbf

```
y_xgb = regr_xgb.predict(X_test)
```

```
y_multirf.sum(axis=1)
```

```
array([ 84.1125,  93.69  ,  99.5325,  94.9675,  91.6625, 104.8675,
       110.915 ,  63.83  , 112.715 , 103.985 , 107.21  ,  94.755 ,
       101.76  ,  79.465 , 122.8025, 103.7075, 128.84  ,  88.6   ,
       143.05  , 106.1975, 104.3675, 121.535 , 124.165 ,  87.255 ,
       126.725 ,  93.5975,  92.7725,  89.38  ])
```

```
y_rf.sum(axis=1)
```

```
array([100., 100., 100., 100., 100., 100., 100., 100., 100., 100., 100.,
       100., 100., 100., 100., 100., 100., 100., 100., 100., 100., 100.,
       100., 100., 100., 100., 100., 100.])
```

```
y_multixgb.sum(axis=1)
```

```
array([ 88.08045 , 105.28155 ,  97.99562 , 105.6268  ,  98.4469  ,
       112.20299 , 110.698586,  60.55402 , 103.17296 , 111.4723  ,
        83.388   ,  89.45125 , 105.47575 ,  99.23981 ,  88.71594 ,
       101.94826 , 113.45636 ,  87.23696 , 112.025085, 106.15552 ,
        85.98609 ,  79.45943 , 105.23216 , 107.59666 , 160.80643 ,
       101.99538 ,  98.053535,  85.83247 ], dtype=float32)
```

```
y_xgb.sum(axis=1)
```

```
array([ 82.50985 ,  87.9862  , 111.65272 , 106.89282 ,  87.54537 ,
       118.22887 ,  98.33519 ,  71.22099 ,  96.27917 , 103.17431 ,
        77.73041 , 102.029625, 112.78526 , 107.03798 ,  92.40184 ,
       106.40093 ,  97.09924 ,  95.292725, 104.641426, 105.17516 ,
        89.75321 ,  99.82206 , 102.922005, 101.816956, 159.372   ,
       101.59437 ,  99.99987 ,  97.00639 ], dtype=float32)
```