Optimizing Model Development and Validation Procedures of Partial Least Squares for Spectral Based Prediction of Soil Properties

Nimrod Carmon

Ph.D. Student
The Remote-Sensing Laboratory
Geography and Human Environment
Tel-Aviv University

Advisor: Prof. Eyal Ben-Dor

THE REMOTE SENSING LABORATORIES

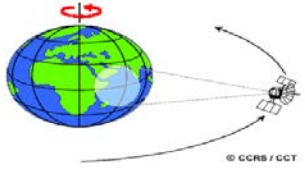
The Porter School of Environmental Studies
בית ספר לימודי הסביבה ע"ש פורטר

Soil Spectroscopy

- **Extracting chemical and physical attributes from spectral data**
- **Execution from various platforms – Handheld, Field, Airborne, Satellite**
- **Endless potential applications – Agriculture, Environmental, Health…**

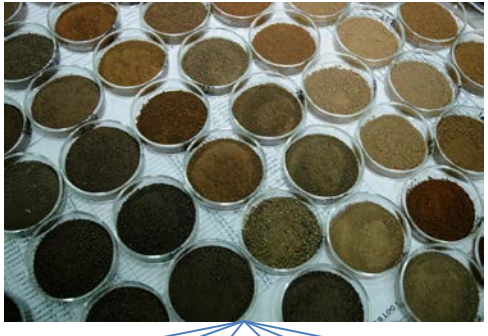




Modeling Soil Properties

Ben-Dor, E., & Banin, A. (1995). **Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties**. *Soil Science Society of America Journal*, *59*(2), 364-372.



SSA
Specific Surface Area

Carbonates
$CaCO_3$

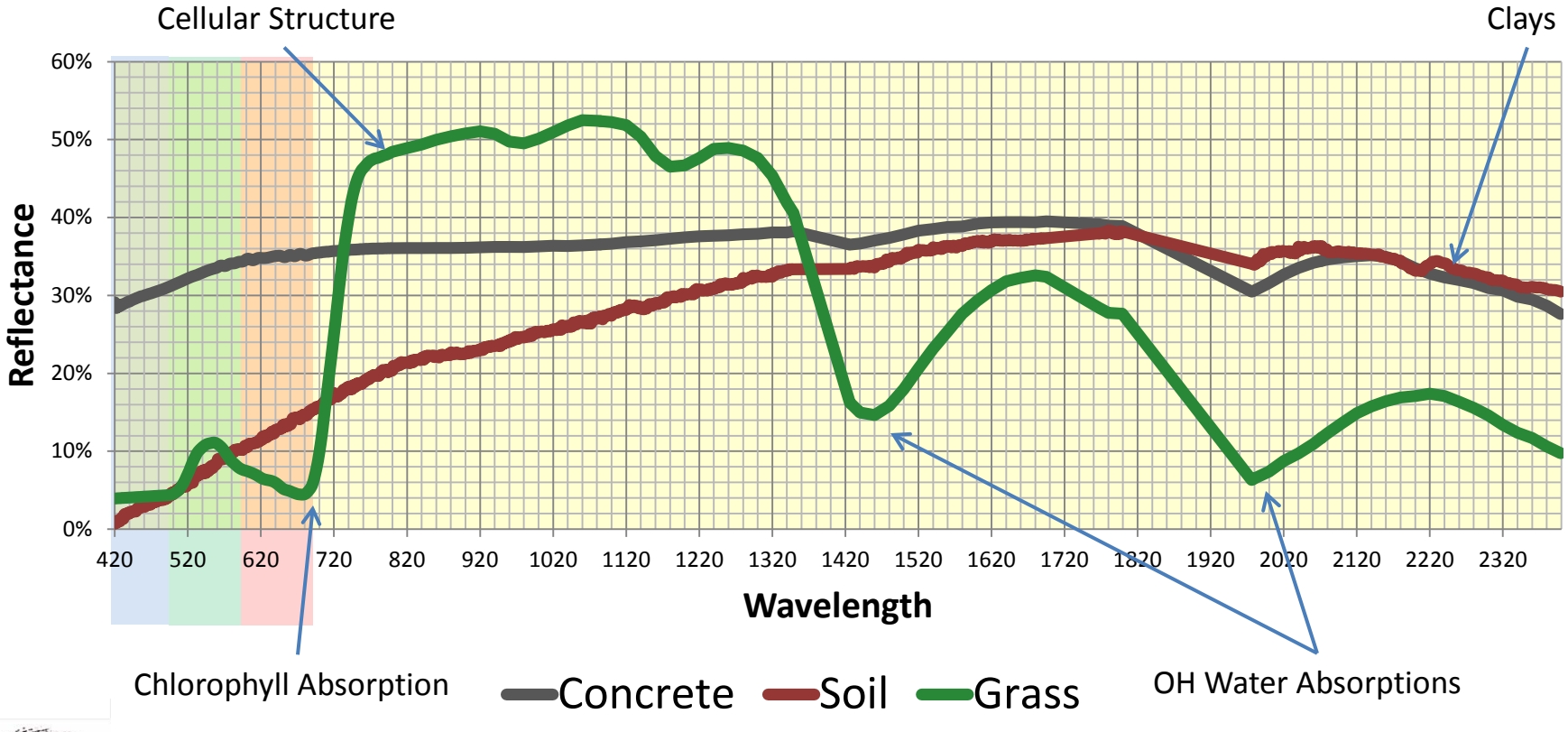
HIGF
Hygroscopic Water

CEC
Cation Exchange Capacity

OM
Organic Matter

Clay
Clay Size Fraction

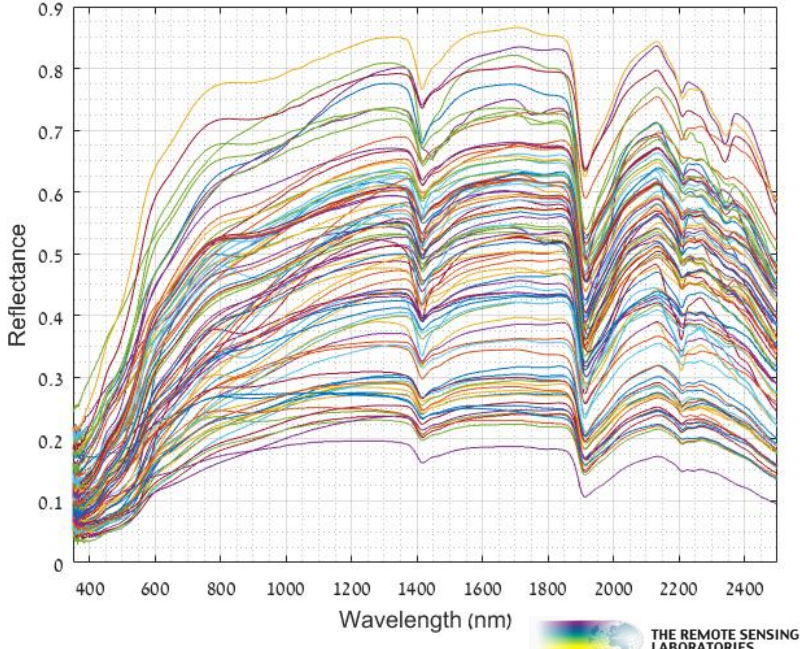
Chemical and Physical Chromophores



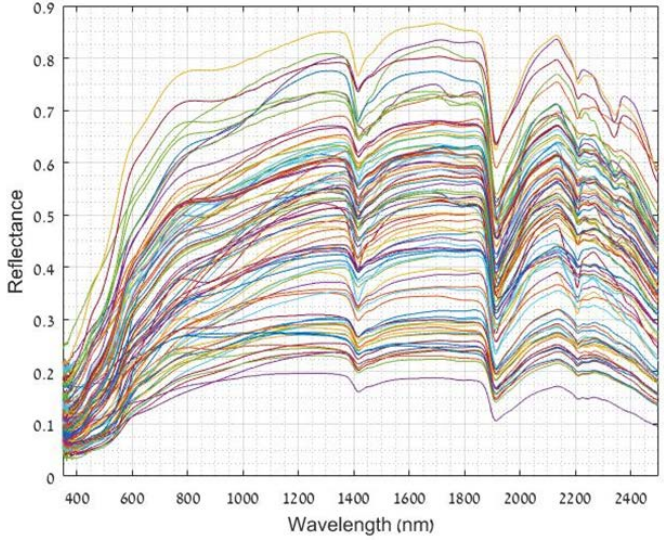
Specific Absorptions Features → Chemistry Composition
Absorption Intensity → Concentration
Overall reflectance → Physical Properties

Big Data Modeling

- **Spectral data is Multivariate – hundreds and thousands of bands**
- **Linear and non-linear spectral ranges**
- **Tens and hundreds of samples**
- **Data from multiple origins**



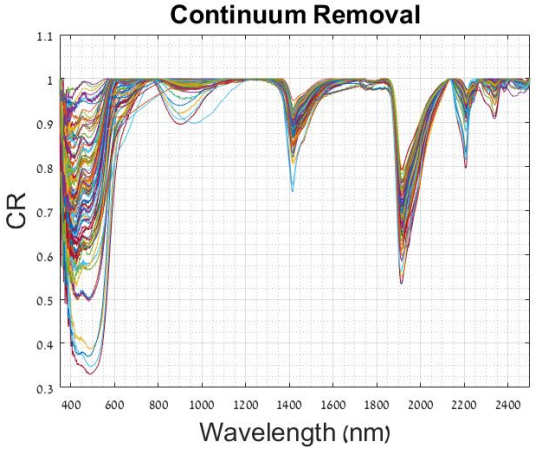
Supervised Machine Learning

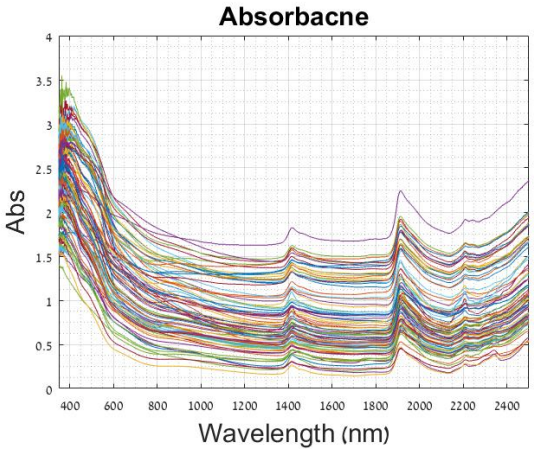


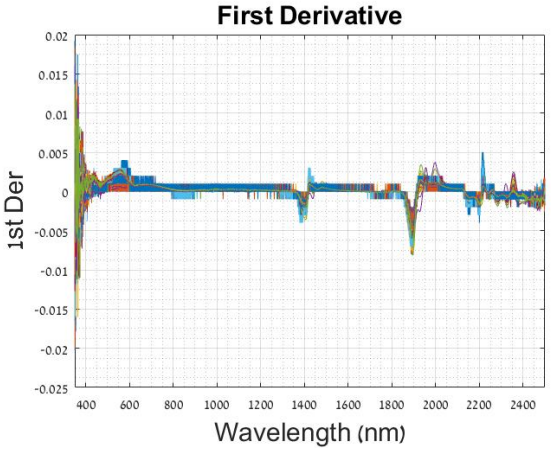
Function Fitting
(MLR, PCA-R, PLS-R, ANN)

Classification
(K-means, SVM, Nearest Neighbor)

Multiple Preprocessing Choices

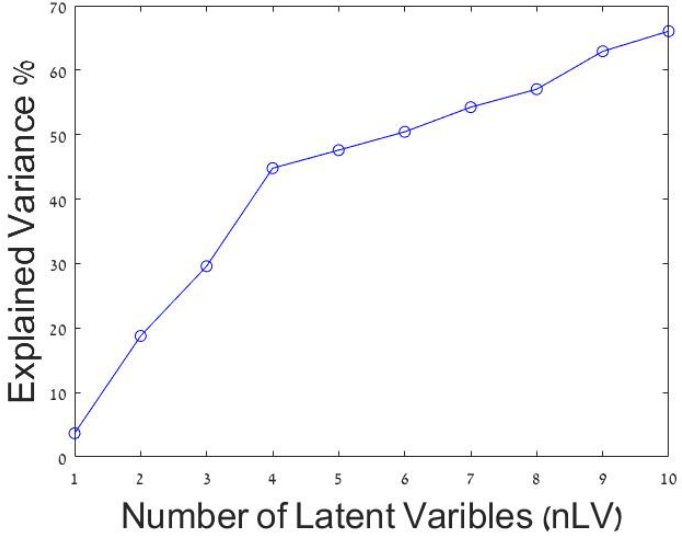






Partial Least Squares – Regression (PLS-R)

- **Dimension rotation in covariance to the modeled properties' values**
- **New factors (latent variables) as predictors with explained variance**
- **Capability to deal with multivariate data (spectra) without overfitting**



How to Apply

- The Unscrumbler©
- SPSS©
- Matlab©
- R

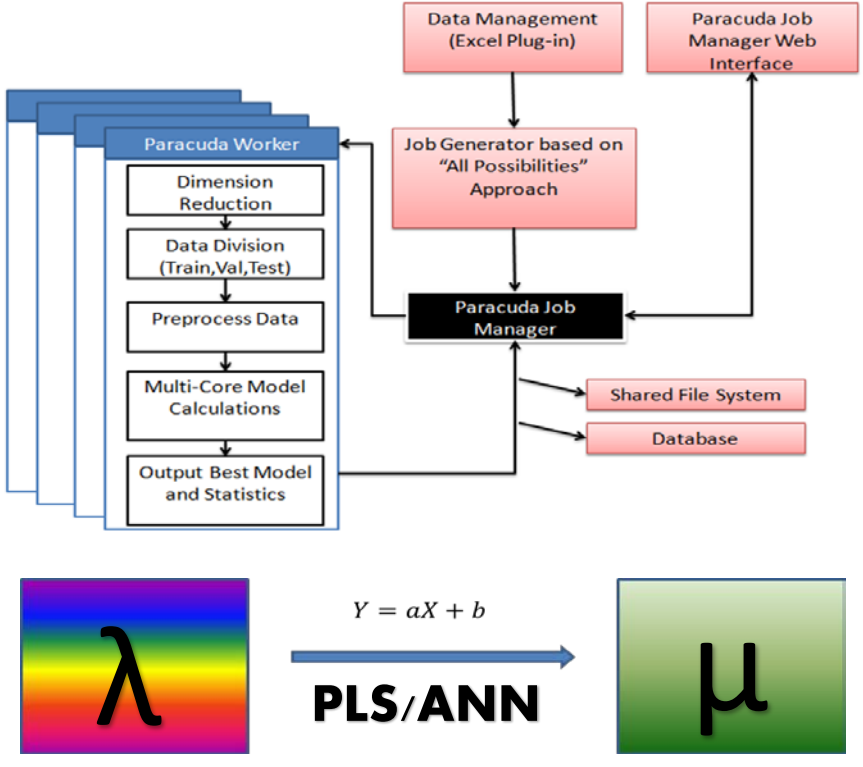
Drawbacks

- Limited output
- Limited configuration
- Requires Programming knowledge
- No Automation for Pre-processing

Solution  **PARACUDA©**

PARACUDA ©

- **PARACUDA Engine**
- **All Possibilities Approach (APA) spectral manipulation**
- **Cloud based**
- **Airborne or Field data modeling**
- **Easy and Automatic utilization**



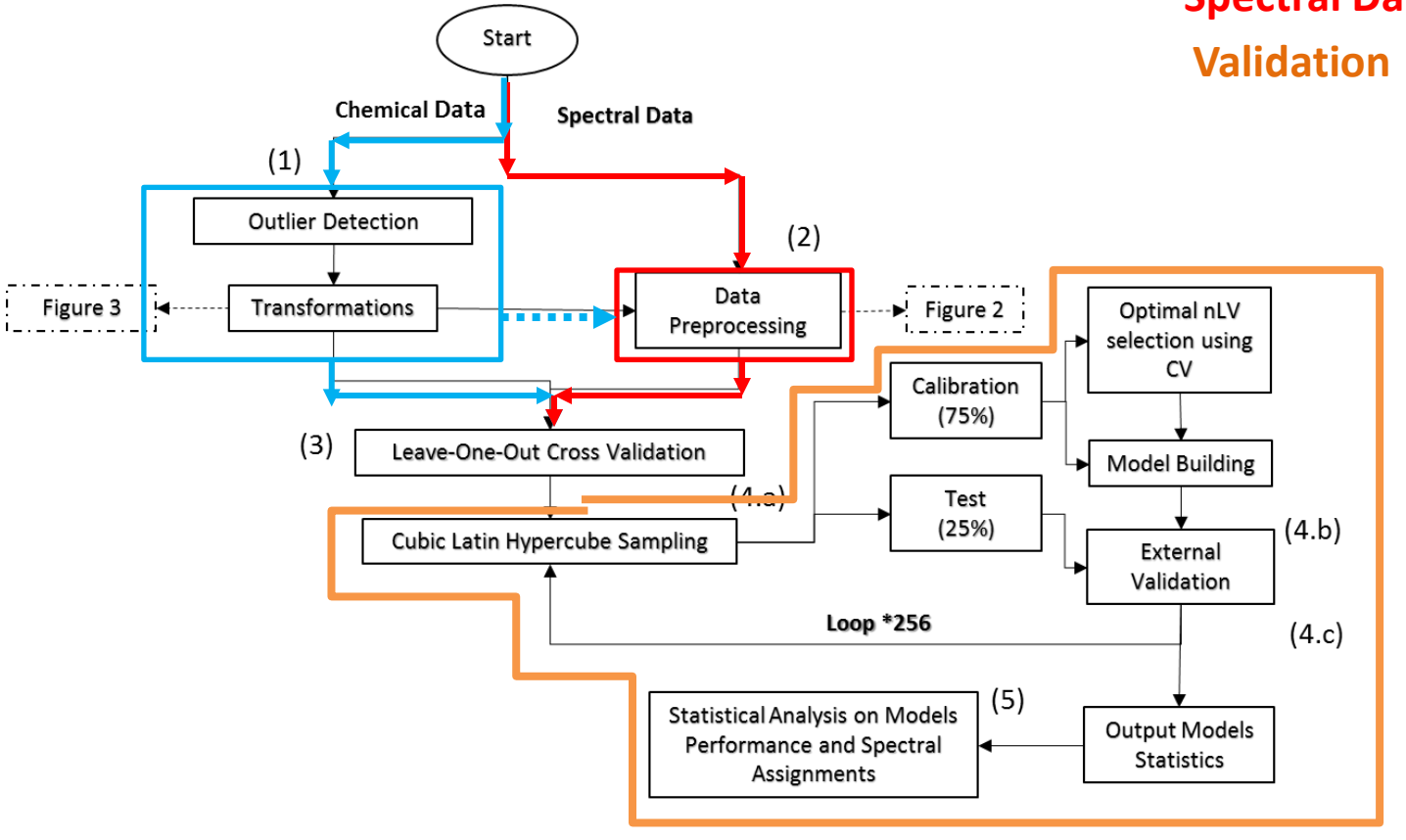
PARACUDA © - the drawbacks

- **ONE model for ONE preprocessing sequence**
- **No spectral assignments**
- **No multi-threading**
- **No transformation to modeled attributes**
- **Limited validation technique**

Solution → **PARACUDA II ©**

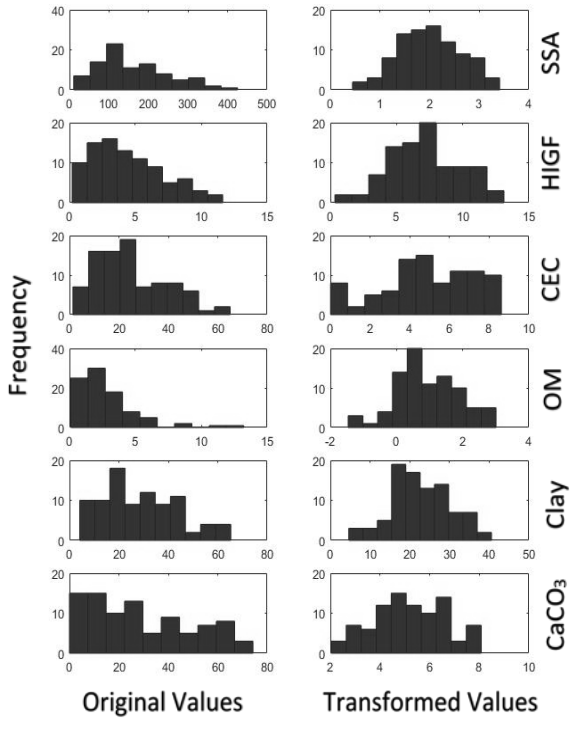
PARACUDA II – Main Structure

Chemical Data
Spectral Data
Validation

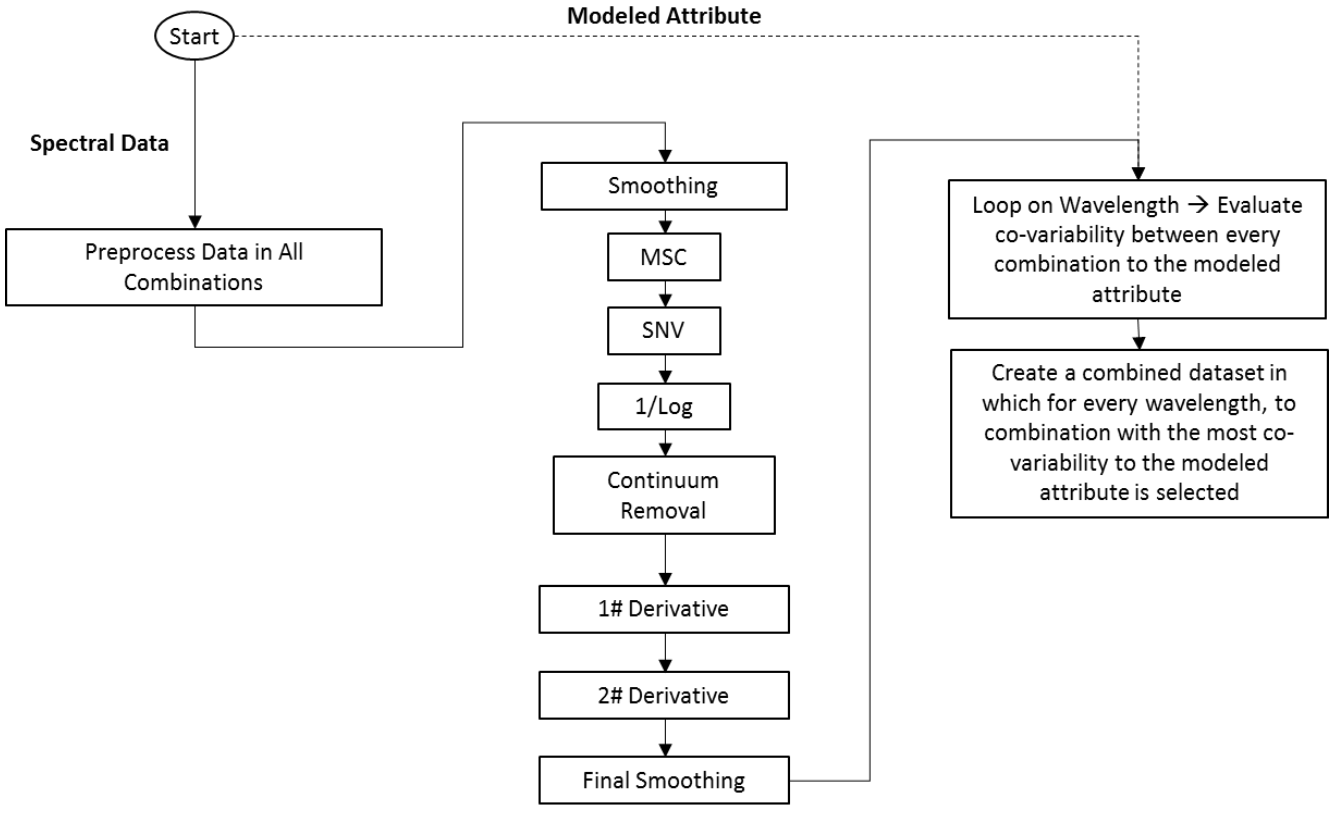


Transformations to modeled values (chemistry)

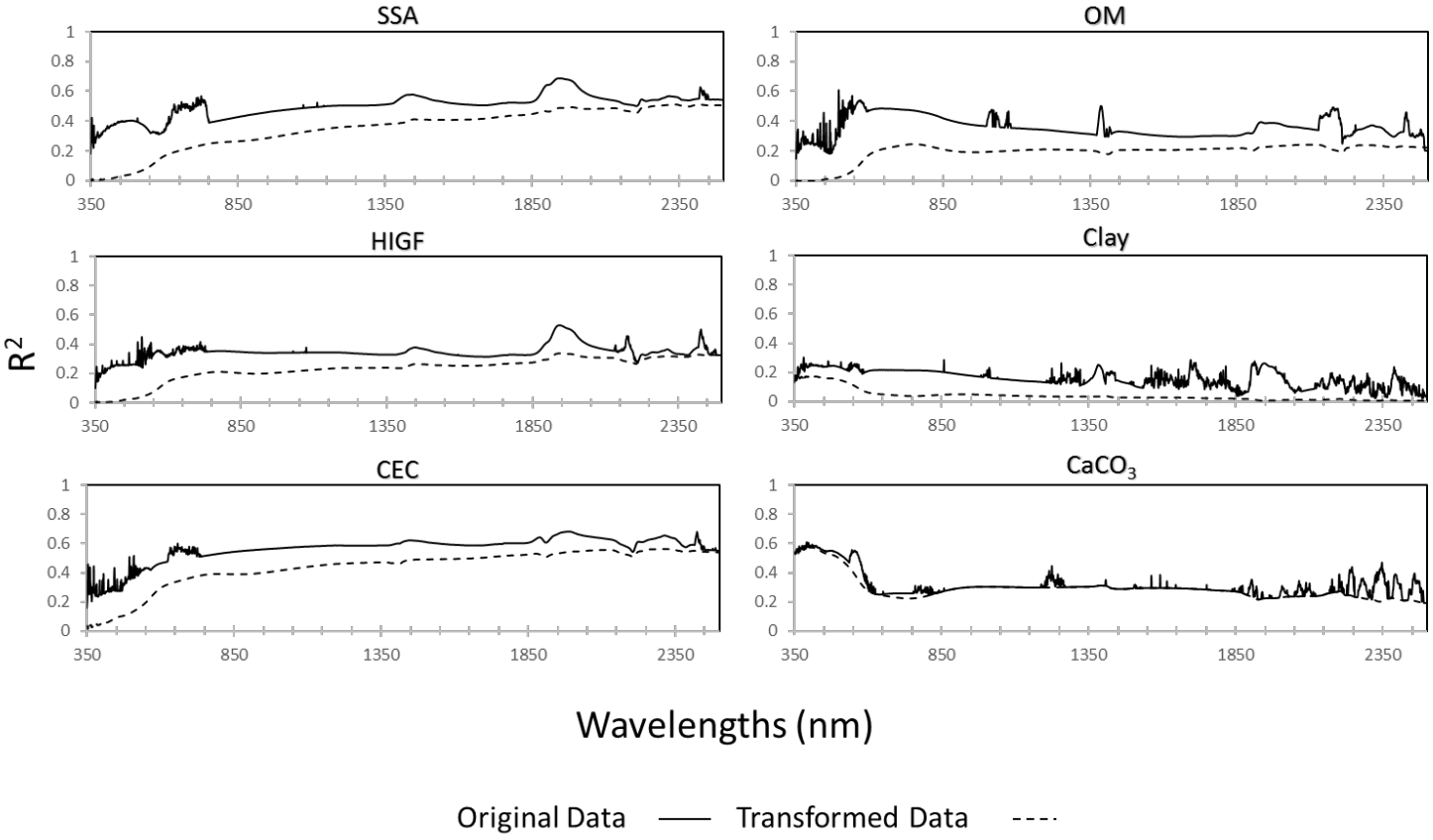
- **Closer to Normal Distribution**
- **Outlier detection**
- **Multiple algorithms (box-cox, Log_x, sqrt)**
- **Normality test**



PARACUDA II – Preprocessing Module



Observing Preprocessing with Correlograms

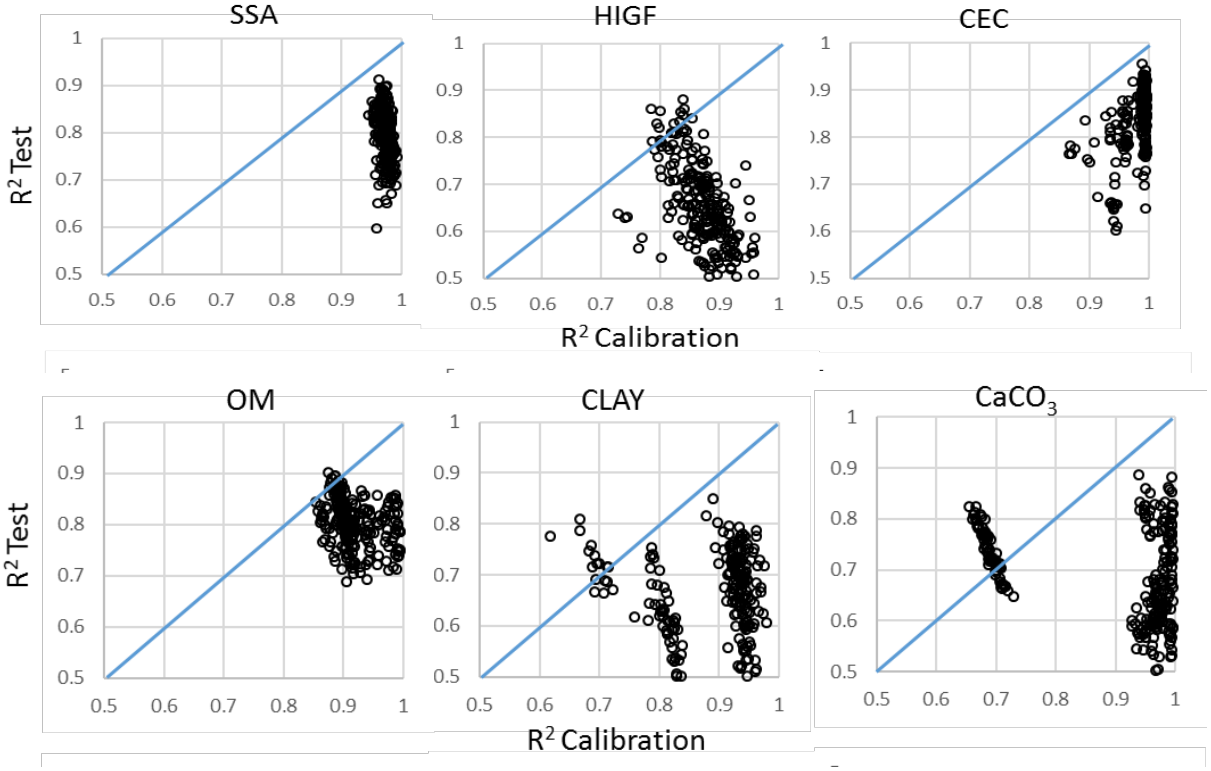


Validation Techniques

- **K-fold / leave-one-out cross validation**
- **Internal validation (75% Calibration, 25% Test)**
- **Internal validation * 256**

Validation 1 – Model Population

- 1:1 line for $R^2_{Cal} > R^2_{Test}$
- **Evaluating model clusters**



Validation 1 – Model Population

- 256 validation iterations
- Analysis of models population
- Reporting both the population performance and best model

Results Comparison

Soil Attribute	Calibration R^2			Test R^2			RPD		SEP	
	Mean	STD	R^{2*}	Mean	STD	R^{2*}	Mean	STD	Mean	STD
SSA	0.973	0.009	0.84	0.797	0.056	0.7	2.332	0.339	3.33	0.47
HIGF	0.872	0.042	0.58	0.651	0.091	0.62	1.705	0.318	0.399	0.064
CEC	0.98	0.025	0.82	0.84	0.07	0.64	2.607	0.522	1.077	0.231
OM	0.918	0.036	0.69	0.804	0.045	0.55	2.488	0.328	0.406	0.054
Clay	0.889	0.077	0.76	0.63	0.119	0.56	1.577	0.234	0.939	0.146
$CaCO_3$	0.894	0.129	0.7	0.692	0.095	0.69	1.67	0.367	13.215	2.86

Statistics of 256 models' performance for each modeled soil attribute. R^{2*} is performance reported by Ben-Dor and Banin (1995).

Validation 2 – Best Available Model

- **Best available model out of 256 iterations**
- **Capability to apply on any given data (point and image)**

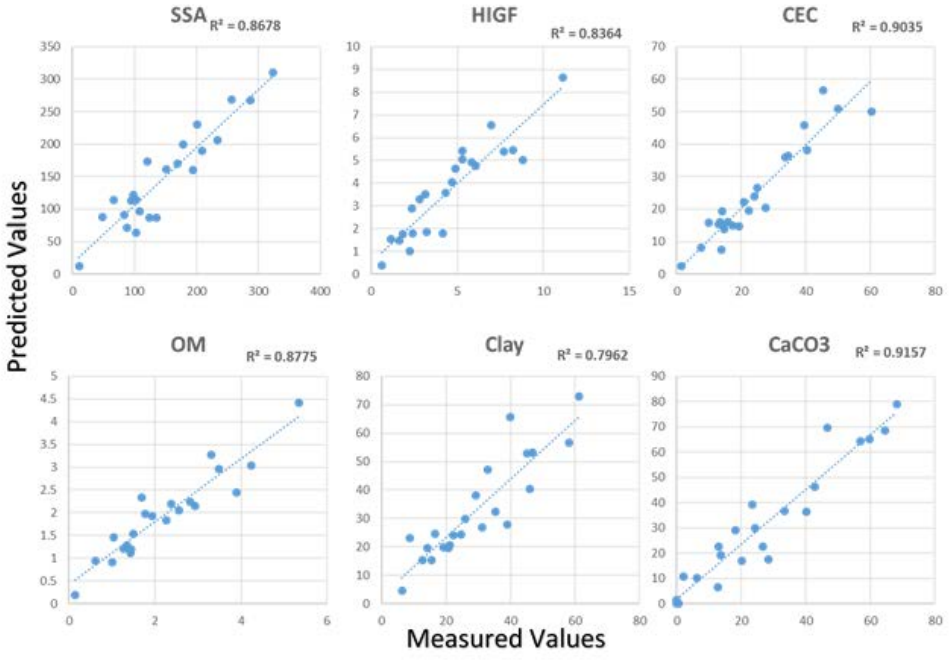


Fig. 8. Best models' prediction performance.

Validation 2 – Best Available Model

- **Extremely high performance**
- **High significance**
- **No overfitting**

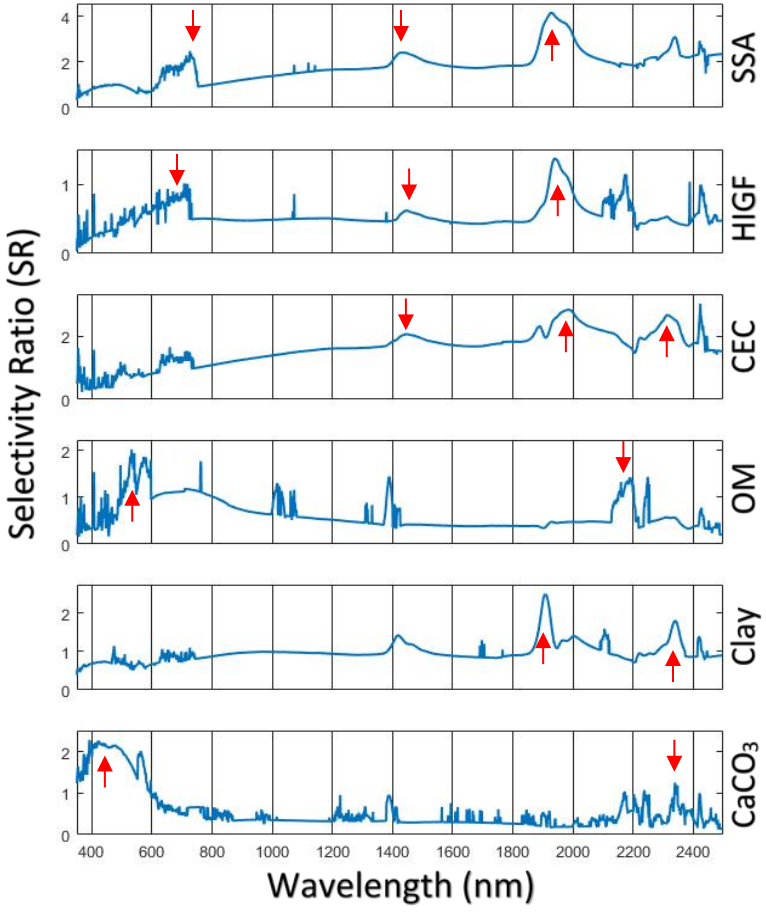
Results Comparison

	R^{2A}	R^{2B}	RPD^{A}	SEP^{A}	SEP^{B}
SSA	0.8678	0.7	3.2445	28.5381	50.2
HIGF	0.8364	0.62	2.2656	1.1752	1.55
CEC	0.9035	0.64	2.9929	4.7148	8.46
OM	0.8775	0.55	3.123	0.5149	1.34
Clay	0.7962	0.56	1.912	7.961	10.3
$CaCO_3$	0.9157	0.69	2.8288	7.4544	11.6

Best available models from PARACUDA II and comparison to Ben-Dor and Banin (1995)

Validation 3 - Spectral Assignments

- **Sensitivity Ratio for every wavelength**
- **Wavelength Importance**
- **Model Evaluation**
- **Bands with high SR → High importance to developed model**



Using PARACUDA II

How to use

- **Create an Excel file with Spectra + Modeled attributes**
- **Run PARACUDA II (one click operation)**

Operation time

- **~ 5 minutes / modeled property (with 100 ASD samples)**

Who can use

- **Anyone**

Samples

	Soil	E1	E2	E3
2	Clay	13.9719	8.8638	25.5598
3	Silt	0	2.5181	3.8851
4	Sand	86.0281	88.618	70.5552
5	SSA	48.5	31.8	102.5
6	HIGF	1.12	0.67	2.15
7	OM	1.27	0.47	1
8	OC	0.59	0.17	0.33
9	CaCO3	1.83	0	0
10	CEC	8.35	4.53	14.18
11	350	0.0558	0.0656	0.0473
12	351	0.0643	0.0563	0.0586
13	352	0.0594	0.0584	0.0561
14	353	0.0566	0.0578	0.0555
15	354	0.0586	0.0517	0.0595
16	355	0.0546	0.0594	0.0535
17	356	0.053	0.0592	0.0499
18	357	0.0535	0.0545	0.048
19	358	0.0575	0.0547	0.0467
20	359	0.0619	0.0524	0.0537
21	360	0.0629	0.052	0.0581
22	361	0.0554	0.0584	0.0489
23	362	0.0504	0.0595	0.0451
24	363	0.048	0.0567	0.0431

Modeled Data

Spectral Data

Conclusions

- Extremely fast operation
- Parallel programming (multi-threading + GPU)
- Statistical transformations to modeled attributes
- Sophisticated APA preprocessing
- Internal iterative validation
- New statistics
- Spectral assignments
- Best available model
- One-click operation

Questions?

THE REMOTE SENSING LABORATORIES

The Porter School of Environmental Studies
בית"ס לימודי הסביבה ע"ש פורטר

Email: Nimrod.RSLab@gmail.com