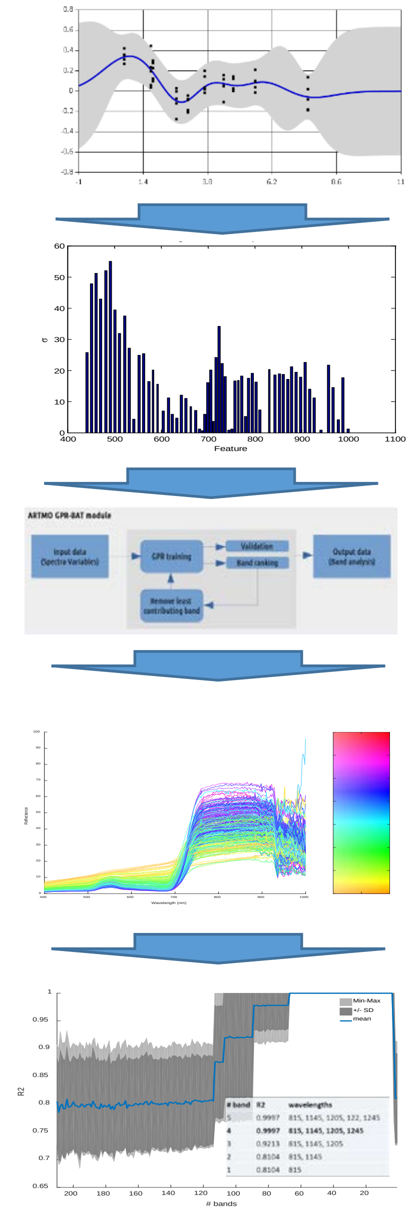
Automated spectral band selection for optimized vegetation properties retrieval using Gaussian processes regression

Jochem Verrelst, Juan Pablo Rivera, Anatoly Gitelson, Jesus Delegido,
Shari Van Wittenberghe, José Moreno, Gustau Camps-Valls
Image Processing Laboratory, Univ. of Valencia (Spain)

EARSeL Imaging Spectroscopy Workshop
19 April2017

Background

- **Rationale: need for band selection**
- **Gaussian processes regression (GPR)**
- **GPR band analysis tool (GPR-BAT)**
- **Leaf/canopy *R* & *SIF* datasets**
- **Automated band analysis**
- **Conclusions**

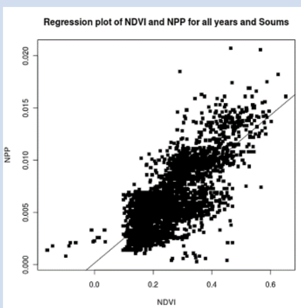


Rationale vegetation properties mapping

Parametric regression	Nonparametric regression	RTM inversion
Spectral relationships that are sensitive to specific vegetation properties	Advanced techniques that search for relationships between spectral data and biophysical variables	Models that simulate interactions between vegetation and radiation

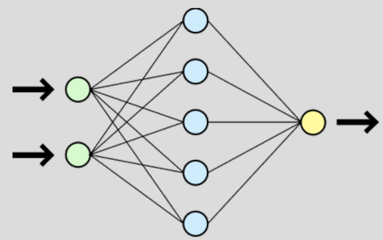
Parametric regression

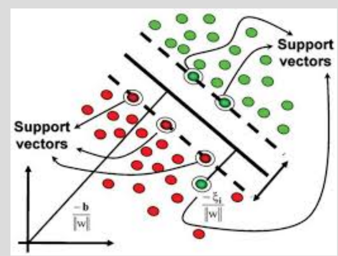
$$NDVI = \frac{(\rho_{NIR} - \rho_{RED})}{(\rho_{NIR} + \rho_{RED})}$$



typically 2 to 4 bands

Nonparametric regression

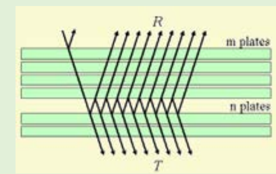


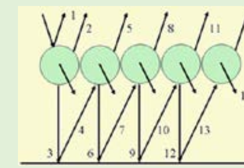


Often all ands

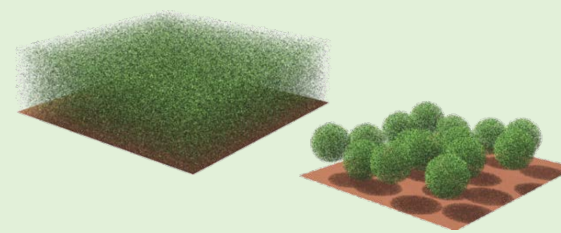
RTM inversion

leaf





canopy



Often all ands

Variable-driven methods

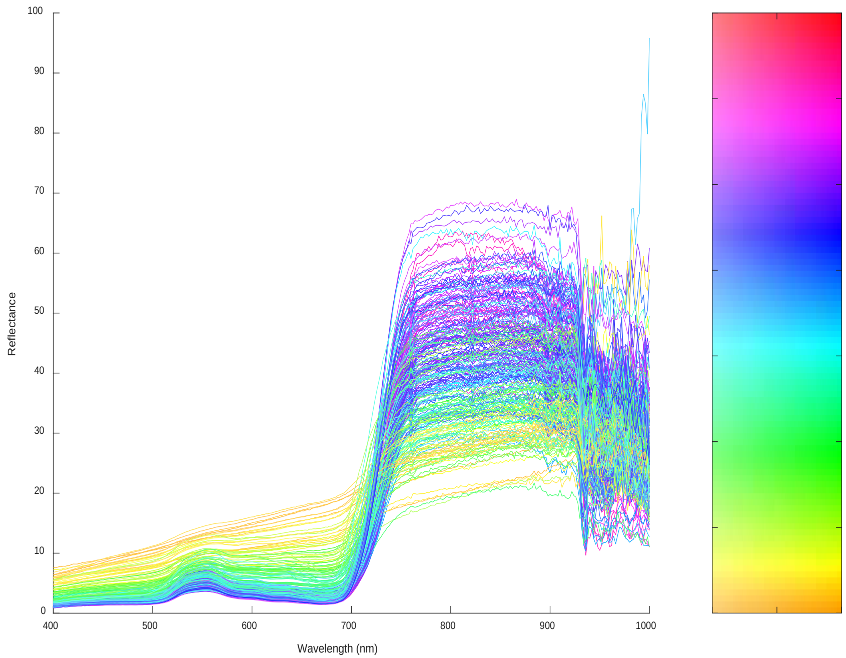
- **data-driven**
- Non-parametric regression more powerful than parametric regression

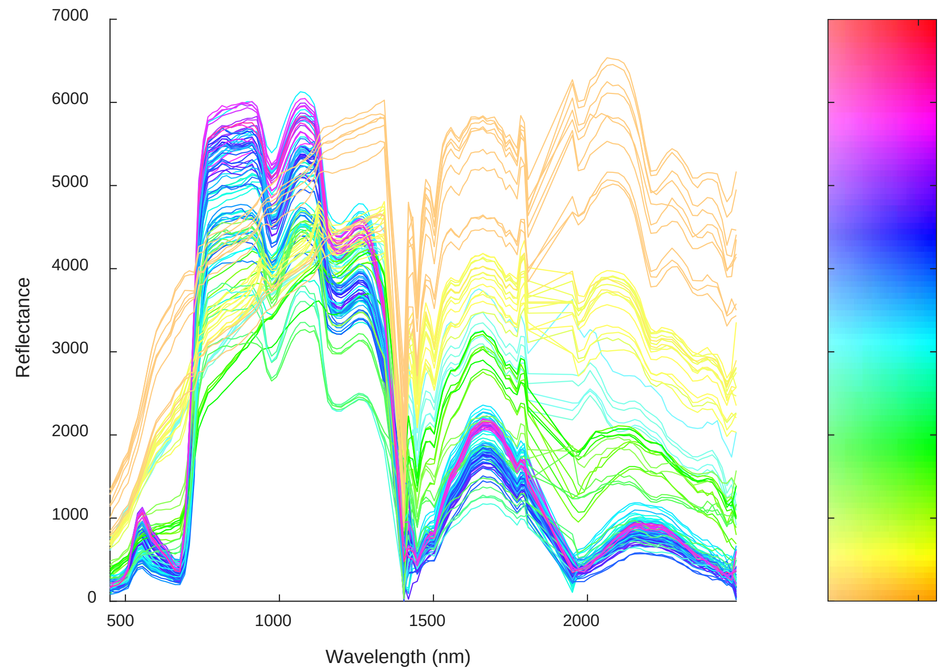
Radiometric methods

- **Spectral fitting**

In regression, band selection is almost a mandatory step when using spectroscopy data.

Which bands to select?

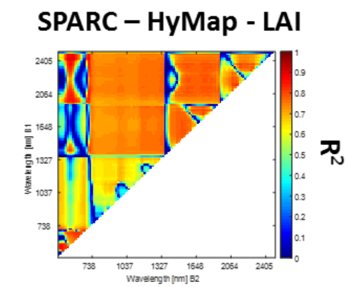




- Using all bands is not recommended
- Using existing vegetation indices (VIs) is questionable

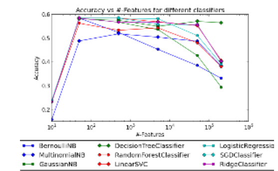
Since each dataset is different, there is a need for an automated band selection method.

For regression mapping applications, current band selection methods are tedious and incomplete



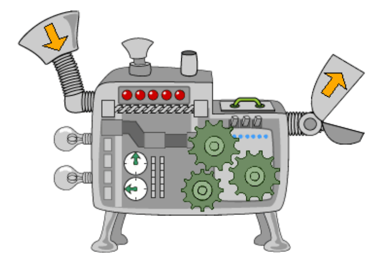
Band analysis methods applied to VIs: systematically analyzing all possible band combinations.

☹ **Limitations**: -Tedious
- Restrict to combinations of 2 or at most 3 bands only



☹ Various band optimization methods developed in classification but **they do not provide spectral information**.

A user-friendly tool is missing that automatically provides the relevant bands for predicting continuous variables (regression).

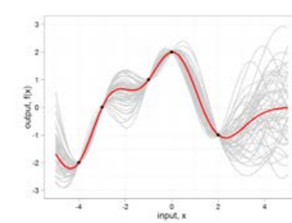


<u>Required:</u>

- ✓ **Identifies minimum number of bands needed for acceptable results**
- ✓ **Gives optimal number of bands**
- ✓ **Gives spectral location of bands**

Some nonparametric methods provide band relevance info (Feilhauer et al., 2015), but none implemented into a ready-to-use tool.

The machine learning method **Gaussian processes regression (GPR)** seems particularly attractive: ☺

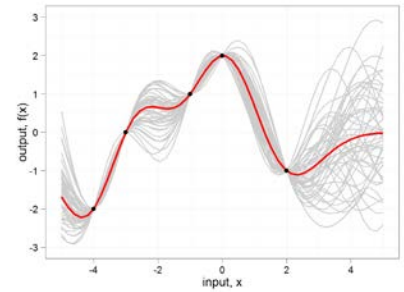


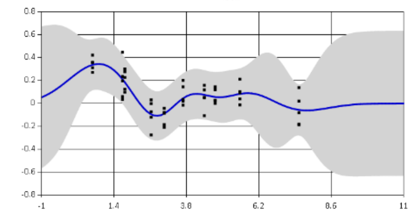
Gaussian Processes Regression (GPR)

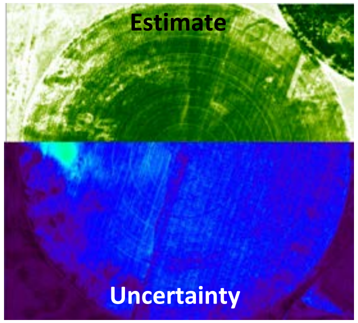
- A **GPR** model is a **probabilistic** (**Bayesian**) model directly in function space, with no intermediate model or model parameters.

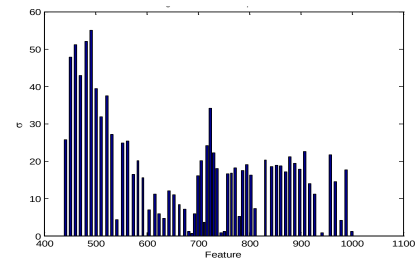
- **GPR** are **equivalent** to **kernel ridge regression**, least square **suport vector machines** (SVM), **Kriging**.

- **GPR alleviates** some **shortcomings** of similar machine learning methods, while maintaining very good numerical performance and stability:
  - GPR is far more **simple than Neural Networks**, and needs **less sample points** ☺
  - **Not only a mean prediction** for each sample (**pixel**), but also **an uncertainty of the prediction** (**confidence interval**). ☺
  - GPR **provide a ranking of features (bands) and samples (spectra)**, thus partly **overcoming the blackbox problem**. ☺
    - http://www.rainsoft.de/projects/gausspro.html



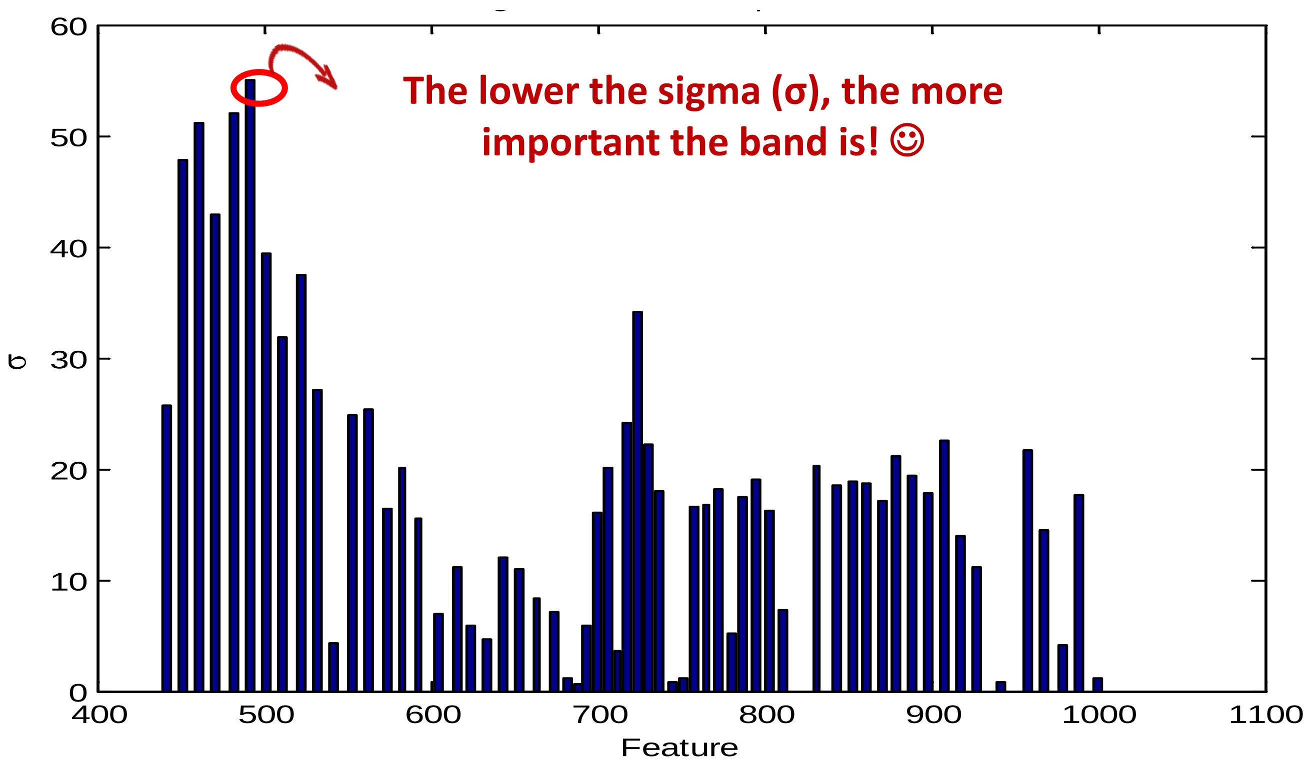






More info on: http://www.gaussianprocess.org/
Rasmussen and C Williams, **Gaussian Processes for Machine Learning**, 2006

The band ranking feature of GPR can used to identify best bands.

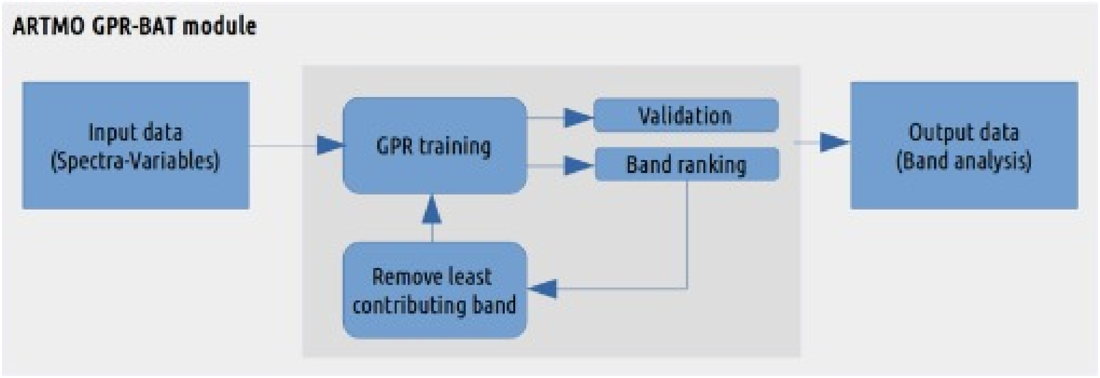


Band ranking results are: (1) data-driven, and (2) for the situation when including all bands.

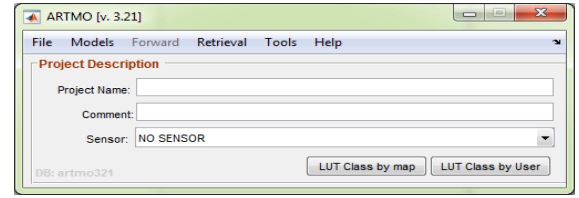
More robust: ***Sequential Backward Band Removal: iteratively removes band with highest sigma (least informative)***

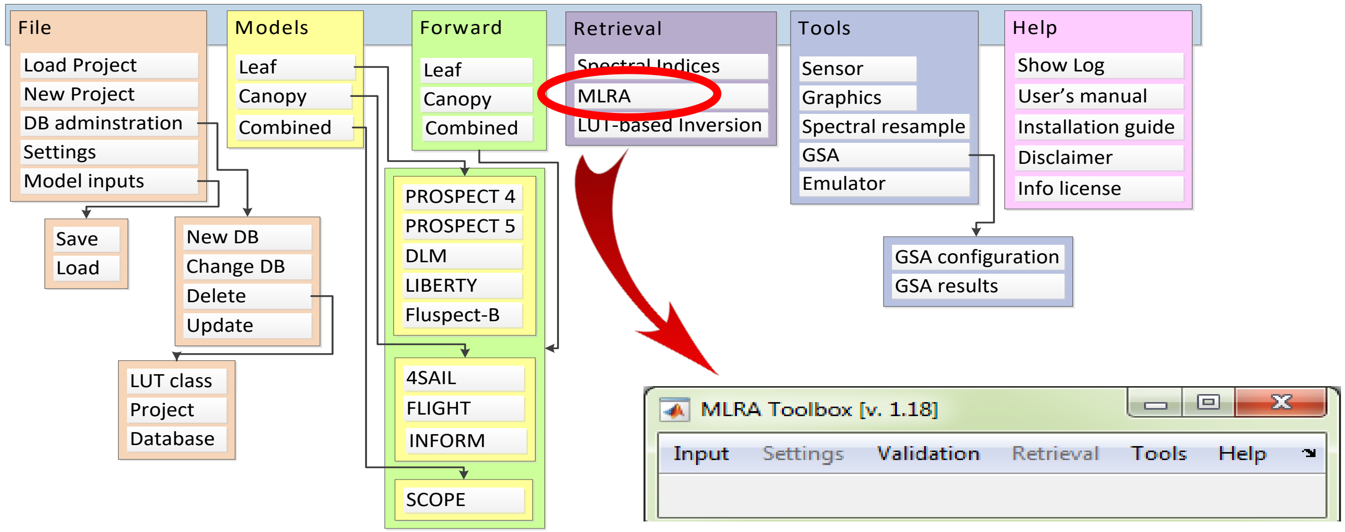
Gaussian processes regression band analysis tool: GPR-BAT

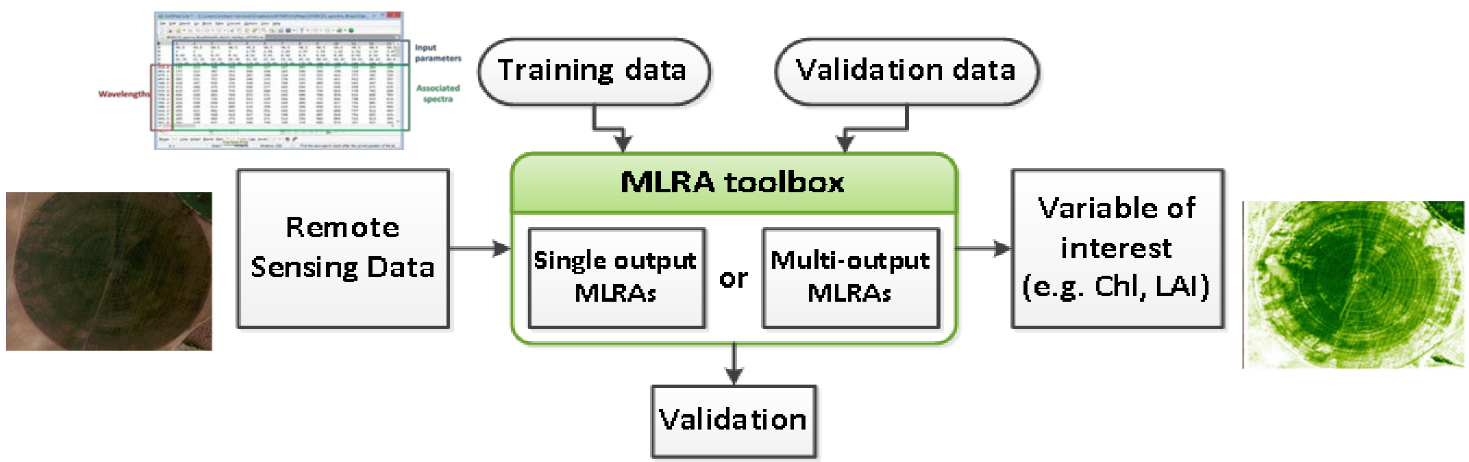
automated



GPR-BAT implemented into a GUI framework: ARTMO

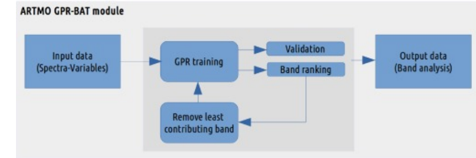




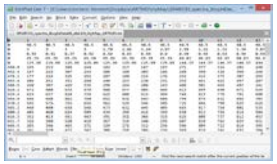


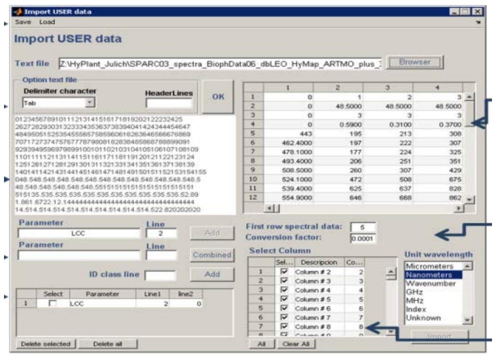
Can be applied to any dataset of spectral data + variable (i.e. not only vegetation)

A few GUIs to click through to run GPR –BAT:

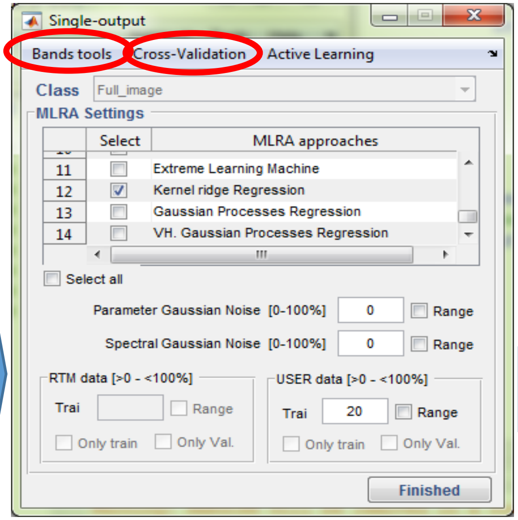


Input



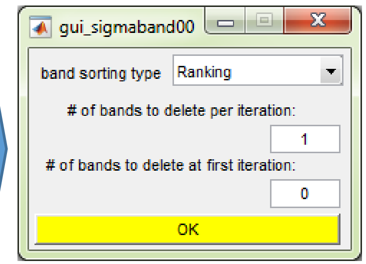


Settings

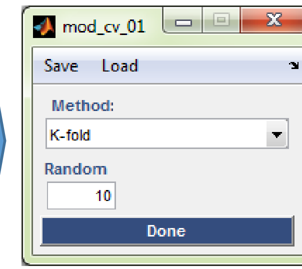


Option of cross-val subsets to make σ ranking more robust (ranking of subsets)

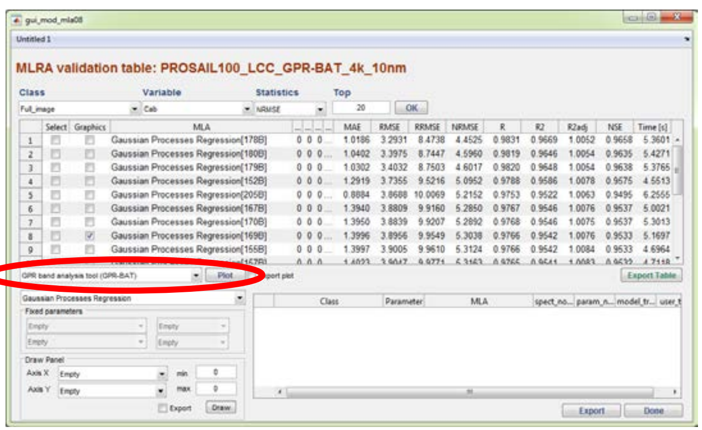
GPR-BAT

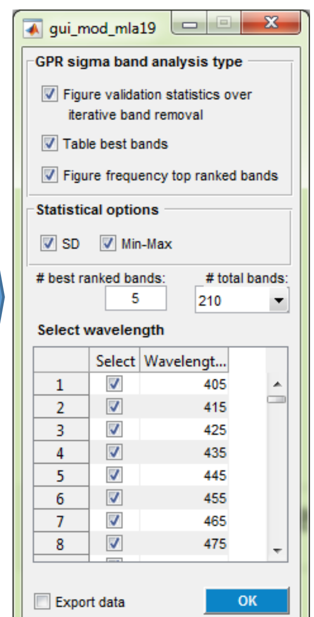


Cross-val

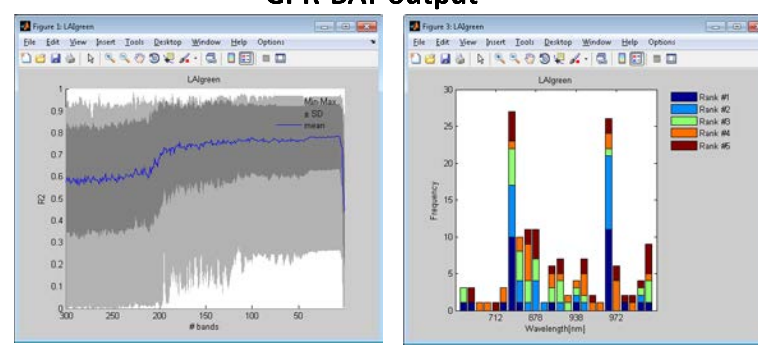


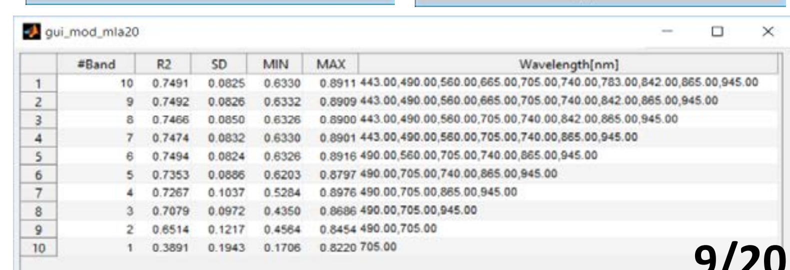
Overview validation





GPR-BAT output



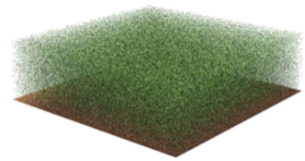
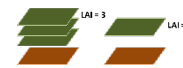


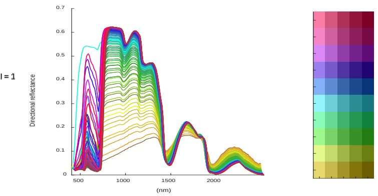
Experiments:

Scale: **Variables:** **Spectra:**

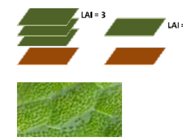
Reflectance (*R*):

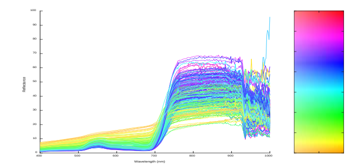
- Simulations:

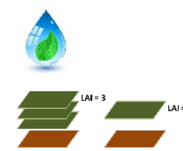


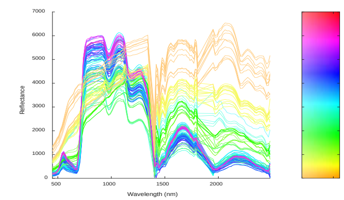
- Field measurements:



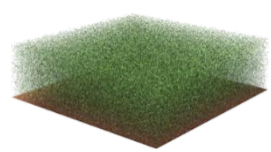
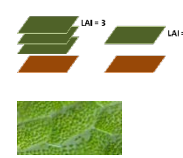
- Airborne measurements:

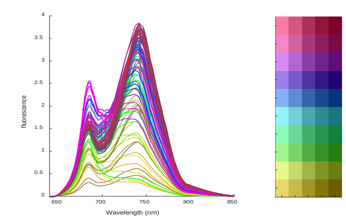





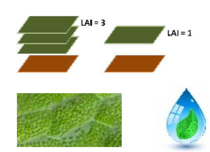
Sun-induced fluorescence (*SIF*):

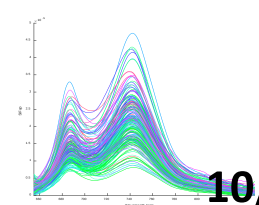
- Simulations:

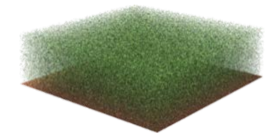


- Leaf measurements:

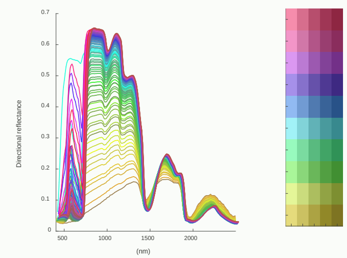


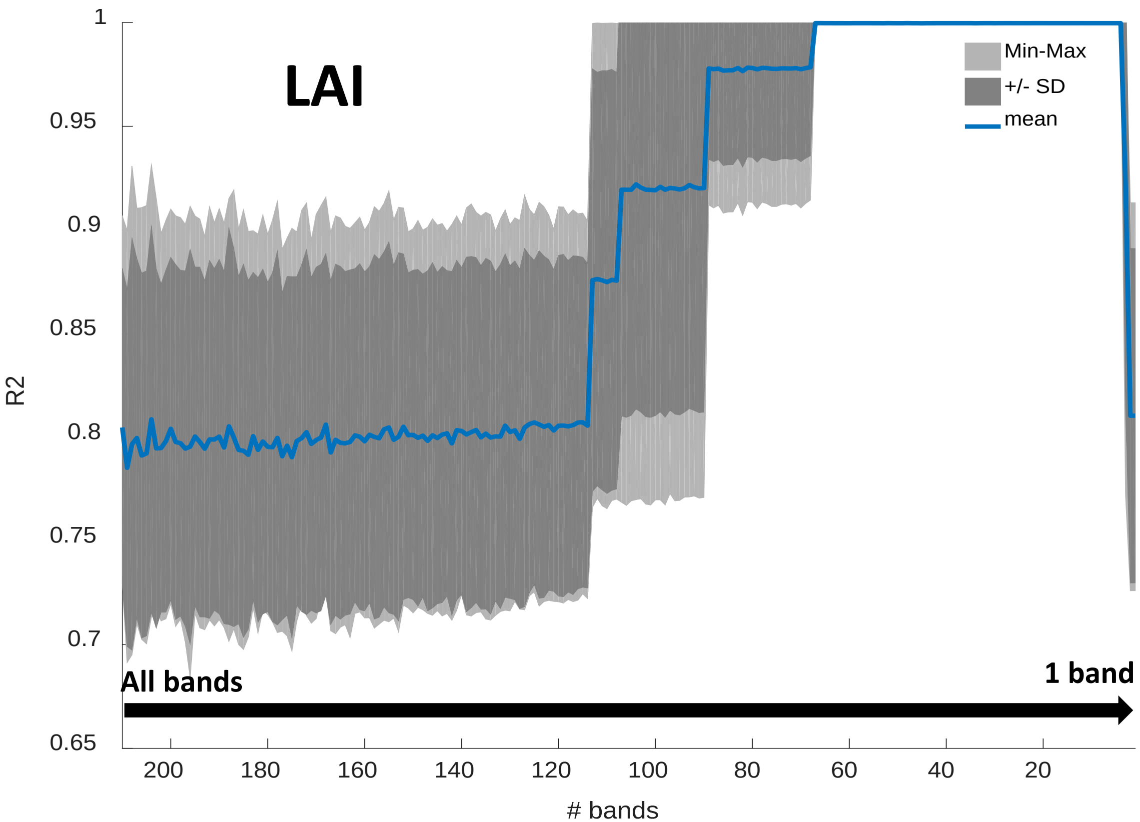
GPR-BAT with simulated data (PROSAIL)



Experimental setup:

- **PROSAIL**: LHS 100#; **Cab, LAI**
- **220 bands@** 10 nm
- **GPR-BAT**: 4-fold CV sampling





# band	R2	wavelengths
5	0.9997	815, 1145, 1205, 122, 1245
4	**0.9997**	**815, 1145, 1205, 1245**
3	0.9213	815, 1145, 1205
2	0.8104	815, 1145
1	0.8104	815

Best performances achieved between 70 and 4 bands.

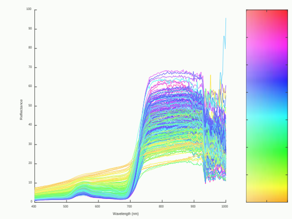
Using all bands or <3 bands not recommended.

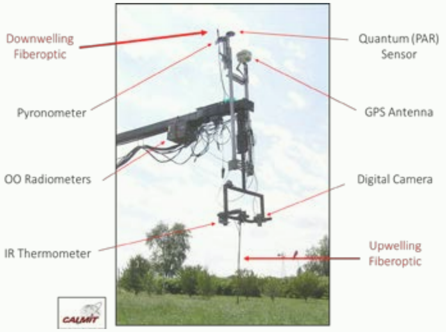
What about real data?

Experimental setup *R* measurements:

<u>ULN test site *(thanks Anatoly ☺)*:</u>

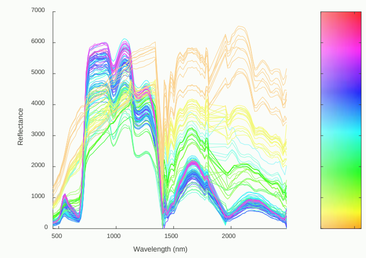
- ~10 years of maize and soya measurements: >260#
- Multiple variables measured, here used: **Cab, gLAI**
- Field spectral data measured by an Ocean Optics: 400-1000 nm
- **301 bands** @2 nm
- GPR-BAT: 10-fold CV sampling



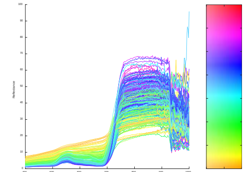


<u>Barrax (Spain) test</u>

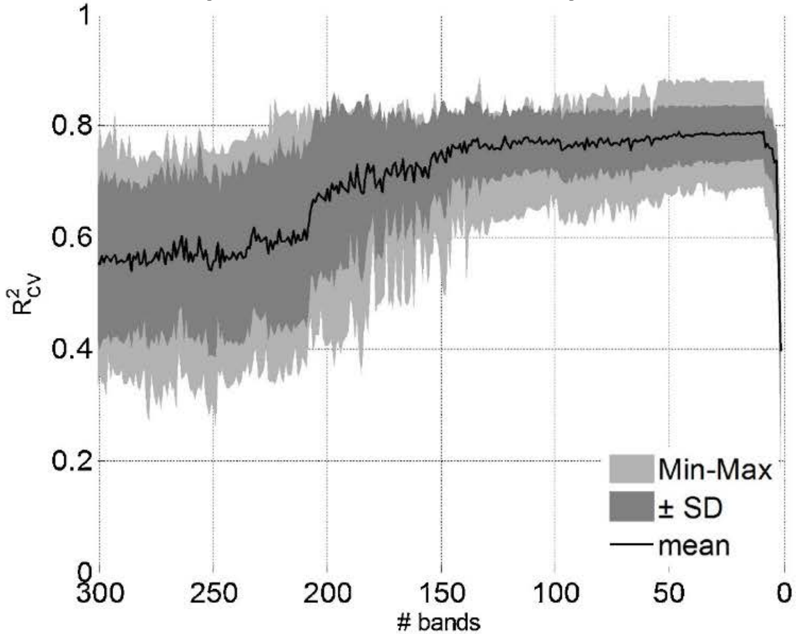
- SPARC dataset (2003/4) over various crops: ~100#
- Multiple variables measured, here used: **CWC, LAI**
- Airborne spectral data measured by HyMap: 450-2500 nm
- **125 bands** at 10-20 nm
- GPR-BAT: 4-fold CV sampling



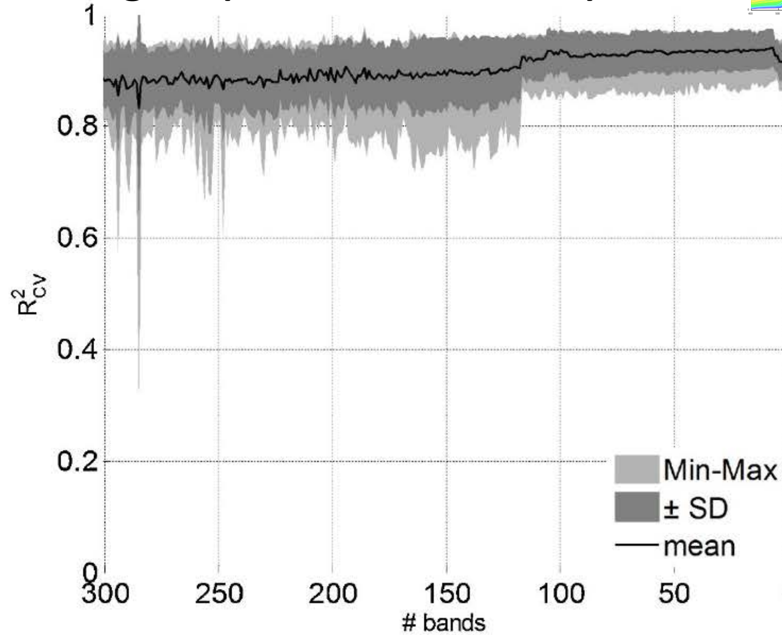
Field data (maize/soybean, OO, 301#b)



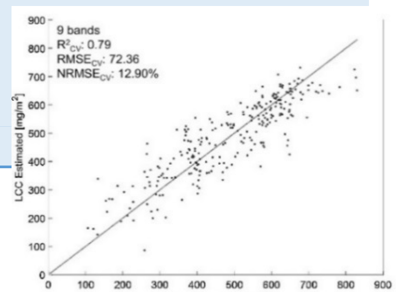
LCC (best with 9 bands)



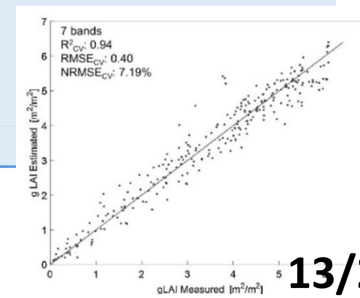
gLAI (best with 7 bands)



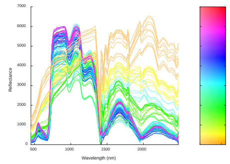
# band	R2	wavelengths
10	0.79	482, 500, 564, 566, 710, 712, 714, 878, 966, 980
9	**0.79**	**482, 500, 564, 710, 712, 714, 878, 966, 980**
8	0.76	482, 500, 564, 710, 712, 714, 878, 966
7	0.77	482, 500, 564, 710, 714, 878, 966
6	0.76	482, 500, 710, 714, 878, 966
5	0.76	500, 710, 714, 878, 966
4	0.73	500, 710, 714, 878
3	0.74	500, 710, 878
2	0.56	500, 710
1	0.40	710



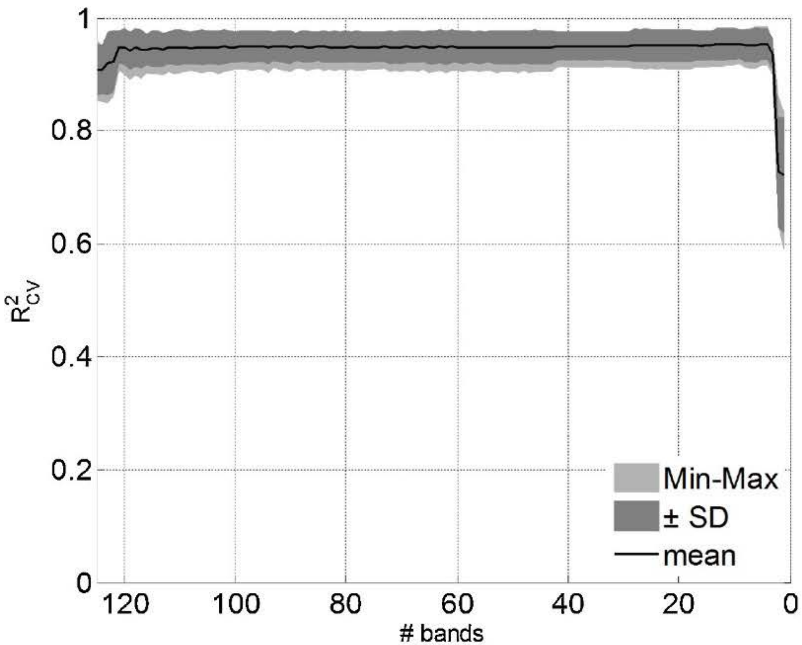
# band	R2	wavelengths
10	0.94	406, 746, 770, 790, 792, 794, 798, 808, 858, 878
9	0.94	406, 746, 790, 792, 794, 798, 808, 858, 878
8	0.94	406, 746, 790, 792, 794, 798, 858, 878
7	**0.94**	**406, 746, 792, 794, 798, 858, 878**
6	0.93	746, 792, 794, 798, 858, 878
5	0.93	746, 792, 794, 798, 878
4	0.91	746, 792, 794, 798
3	0.91	746, 792, 794
2	0.92	746, 792
1	0.64	792



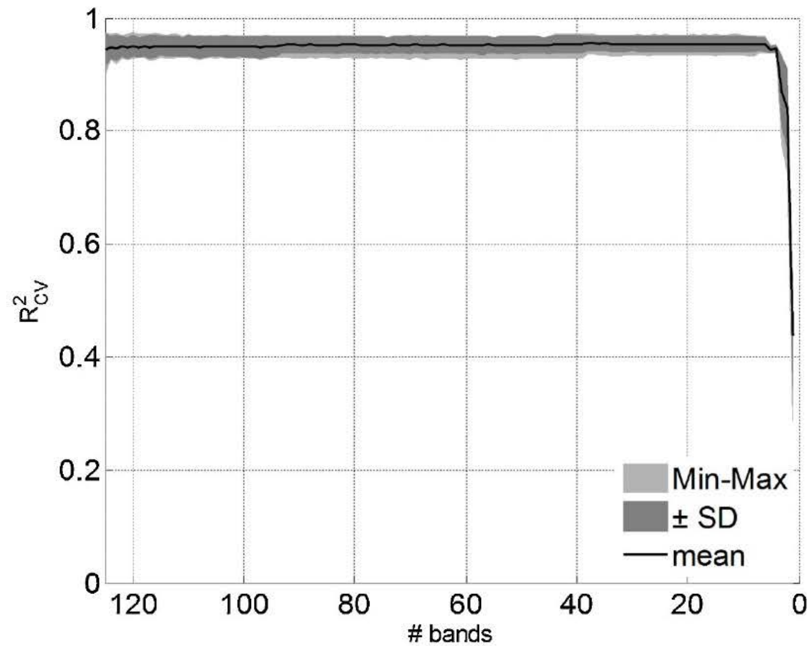
Airborne data (SPARC, Barrax, Spain; Hymap, 125#b)



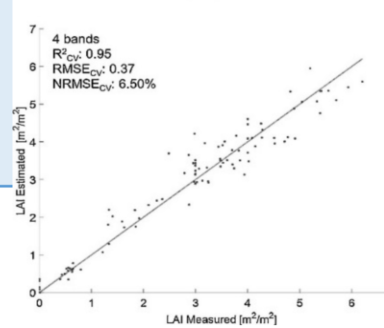
LAI (best with 4 bands)



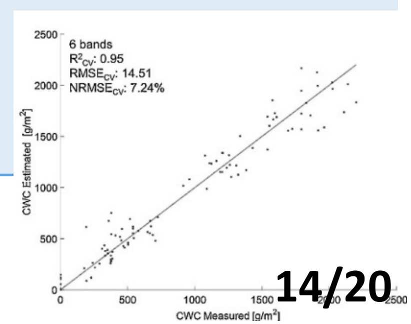
CWC (best with 6 bands)



# band	R2	wavelengths
10	0.95	462, 478, 708, 723, 1215, 1243, 1272, 1327, 1635, 2483
9	0.95	462, 478, 708, 723, 1215, 1243, 1272, 1327, 2483
8	0.95	462, 478, 708, 723, 1215, 1243, 1272, 1327
7	0.95	462, 478, 708, 723, 1215, 1272, 1327
6	0.95	462, 478, 708, 723, 1215, 1327
5	0.95	462, 478, 708, 723, 1327
4	**0.95**	**462, 708, 723, 1327**
3	0.94	462, 708, 1327
2	0.73	462, 1327
1	0.72	462

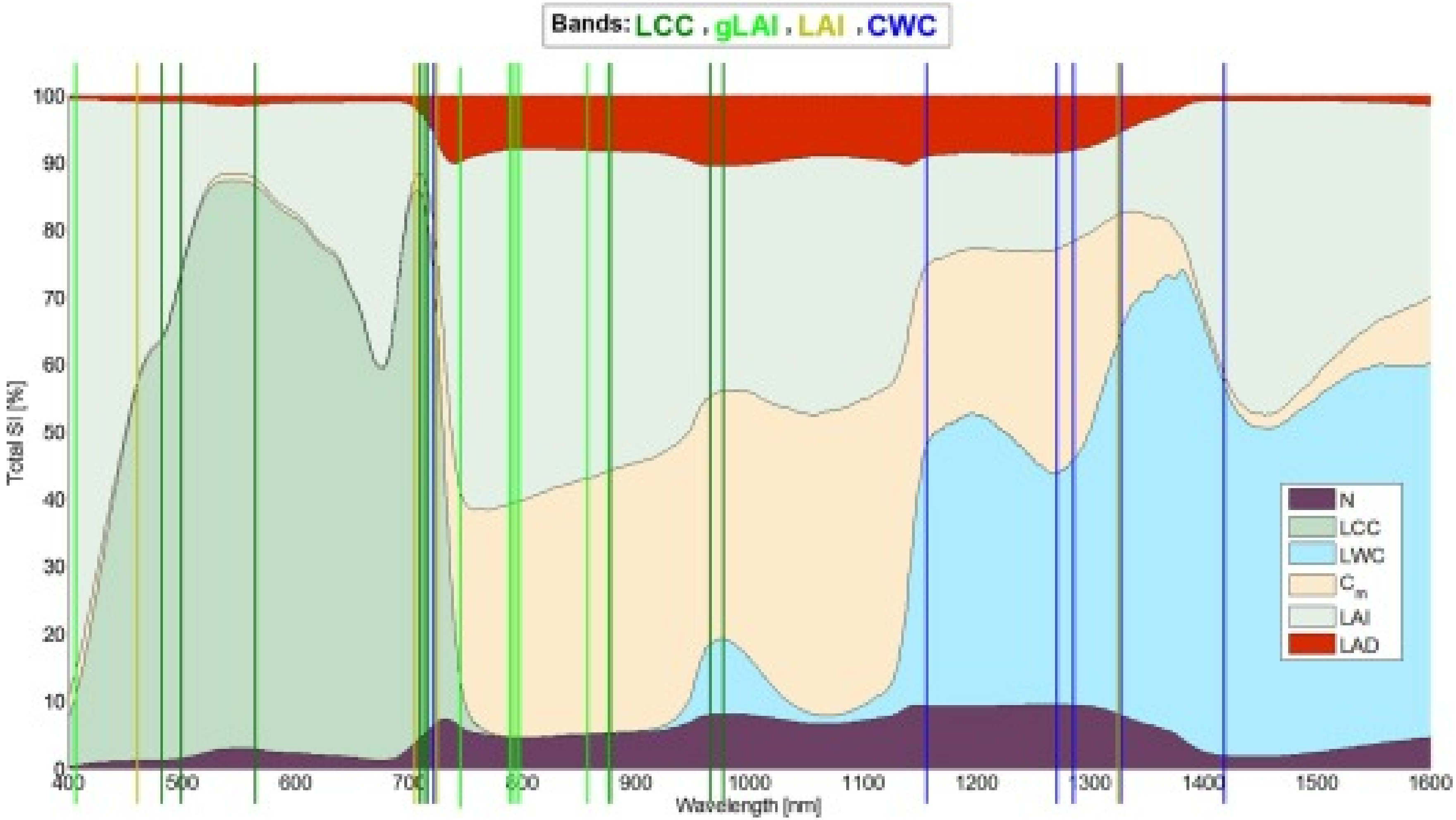


# band	R2	wavelengths
10	0.95	462, 723, 1128, 1157, 1272, 1286, 1299, 1327, 1419, 2483
9	0.95	723, 1128, 1157, 1272, 1286, 1299, 1327, 1419, 2483
8	0.95	723, 1128, 1157, 1272, 1286, 1327, 1419, 2483
7	0.95	723, 1128, 1157, 1272, 1286, 1327, 1419
6	**0.95**	**723, 1157, 1272, 1286, 1327, 1419**
5	0.95	723, 1157, 1272, 1286, 1327
4	0.95	723, 1157, 1272, 1286
3	0.87	1157, 1272, 1286
2	0.84	1157, 1286
1	0.44	1286



Closer look selected bands: comparison with GSA PROSAIL

Best bands for UNL dataset (LCC, gLAI) and SPARC dataset (LAI, CWC) plotted on PROSAIL GSA

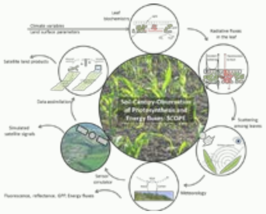


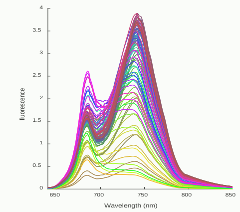
- The band selection of most variables are in agreement with the sensitive regions.
- In case of LCC, there are some secondary bands beyond the LCC region. This can be explained by co-variance relationships.

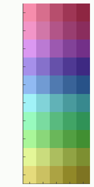
Experimental setup *SIF* measurements:

Experimental setup SCOPE:

- LHS 100# 12 biochemistry/optical variables, e.g. **Cab, LAI**
- **201 bands**: 650-850 nm @1 nm
- GPR-BAT: 4-fold CV sampling

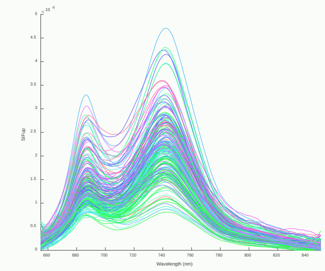


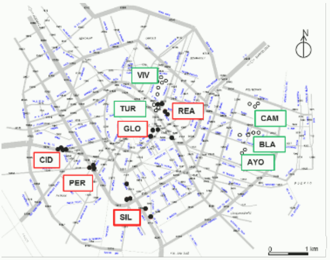




BIOHYPE* dataset: leaf scale SIF measurements:

- 4 urban tree species, >300 leaf spectra of R, T, up/downward SIF
- SIF measurements: **201 bands**: 650-850 nm @1 nm
- Leaf biochemical data:
  - **Specific leaf area (SLA)**
  - **Leaf water content (LWC)**
  - **Leaf Chl content (LCC)**
- GPR-BAT: 4-fold CV sampling

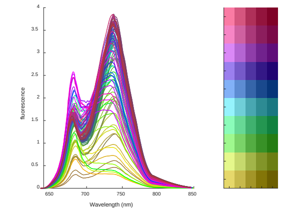




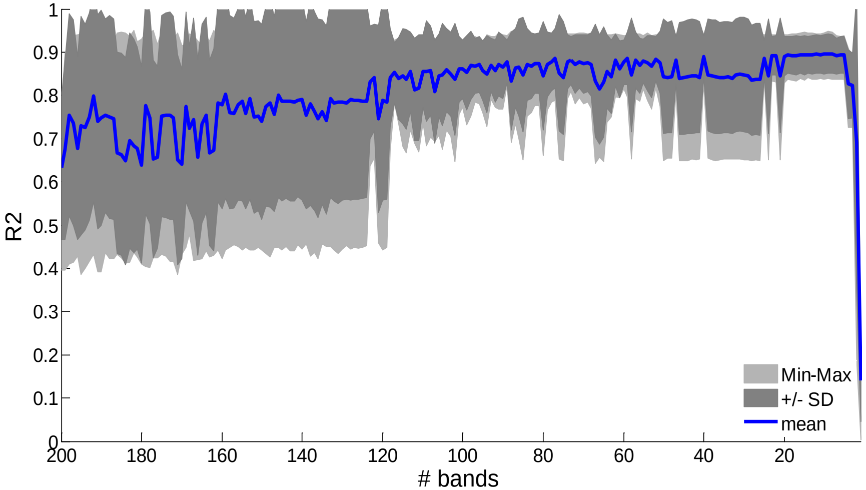
Valencia

*Van Wittenberghe, S., Alonso, L., Verrelst, J., Hermans, I., Delegido, J., Veroustraete, F., Valkce, R., Moreno, J., Samson, R. (2013). Adaxial and abaxial solar-induced chlorophyll fluorescence yield indices of four tree species as indicators of traffic pollution in Valencia. *Environmental Pollution,* 173, p. 29-37.

SCOPE 12 vars, 100#, 4*k*: *SIF* (201#b)

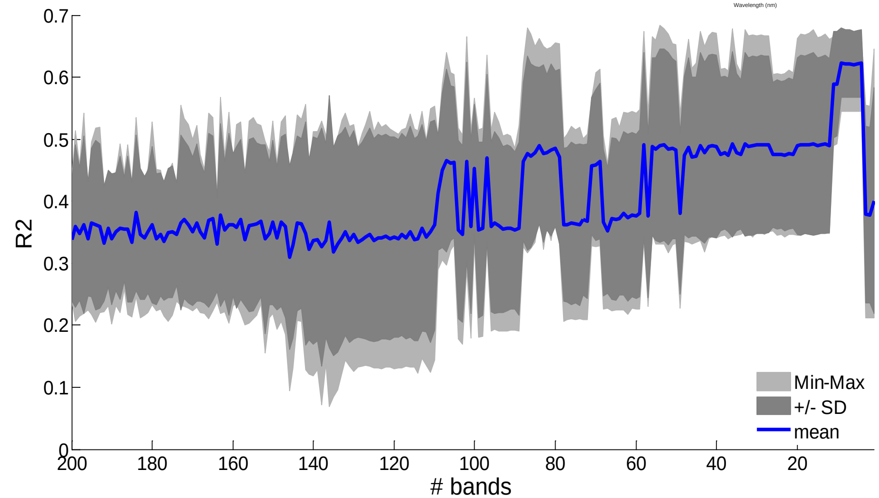


Cab (best at 9 bands)



# band	R2	wavelengths
10	0.89	651, 652, 689, 690, 691 706, 725, 726, 727, 728
9	**0.90**	**651, 652, 689, 690, 691 706, 725, 726, 727**
8	0.89	651, 652, 689, 690, 691 706, 725, 726
7	0.89	651, 652, 690, 691 706, 725, 726
6	0.89	651, 690, 691 706, 725, 726
5	0.89	651, 690, 691 706, 725
4	0.83	651, 690, 691, 725
3	0.82	651, 691, 725
2	0.69	651, 691
1	0.14	691

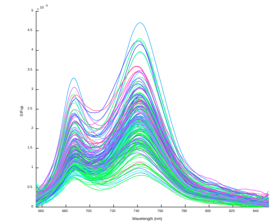
LAI (optimal at 4 bands)



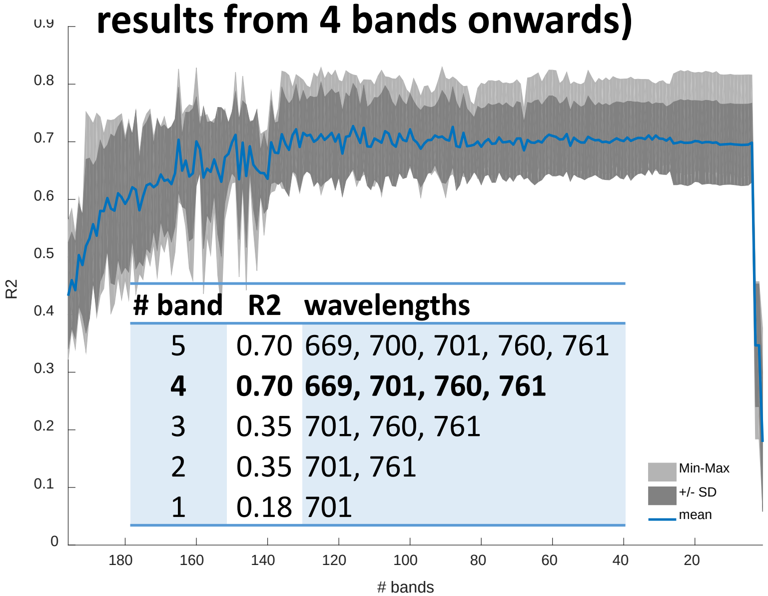
# band	R2	wavelengths
10	0.59	686, 758, 759,760, 765, 766, 767, 768, 795, 796
9	0.62	686, 758, 759, 765, 766, 767, 768, 795, 796
8	0.62	686, 758, 765, 766, 767, 768, 795, 796
7	0.62	686, 765, 766, 767, 768, 795, 796
6	**0.62**	**686, 765, 766, 767, 768, 795**
5	0.62	686, 766, 767, 768, 795
4	0.62	686, 766, 768, 795
3	0.38	766, 768, 795
2	0.38	766, 795
1	0.40	795

- **Same trend as observed as for *R*: suboptimal when using many bands.**
- **Best results with 4-20 bands.**
- **2-bands poor results**

BIOHYPE: upward SIF (200#b)

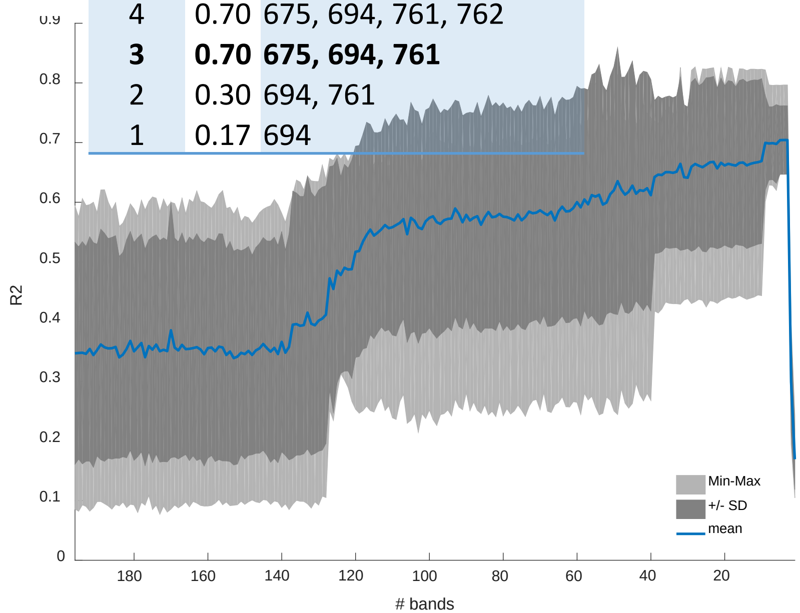


LWC (best with 116 bands, stable results from 4 bands onwards)



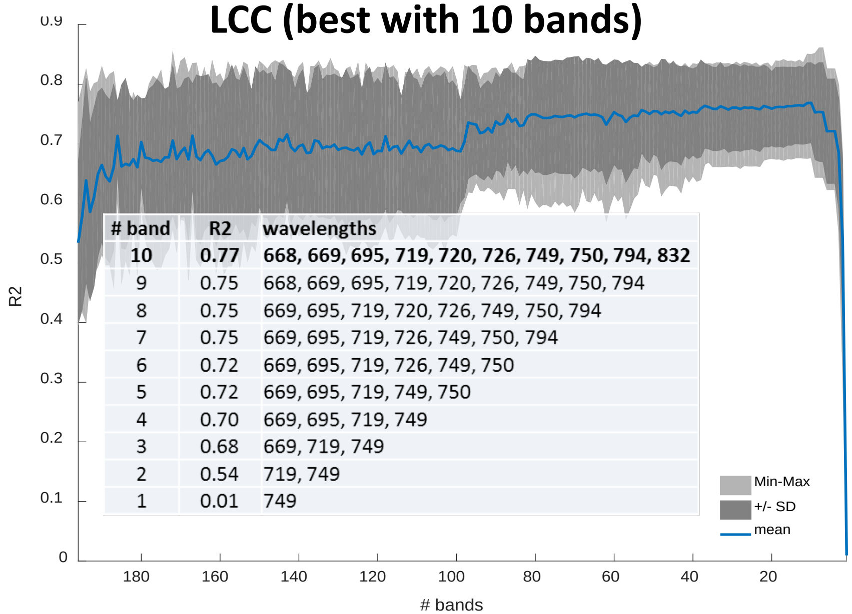
# band	R2	wavelengths
5	0.70	669, 700, 701, 760, 761
4	**0.70**	**669, 701, 760, 761**
3	0.35	701, 760, 761
2	0.35	701, 761
1	0.18	701

SLA (best with 3 bands)



# band	R2	wavelengths
5	0.70	675, 676, 694, 761, 762
4	0.70	675, 694, 761, 762
3	**0.70**	**675, 694, 761**
2	0.30	694, 761
1	0.17	694

LCC (best with 10 bands)



# band	R2	wavelengths
10	**0.77**	**668, 669, 695, 719, 720, 726, 749, 750, 794, 832**
9	0.75	668, 669, 695, 719, 720, 726, 749, 750, 794
8	0.75	669, 695, 719, 720, 726, 749, 750, 794
7	0.75	669, 695, 719, 726, 749, 750, 794
6	0.72	669, 695, 719, 726, 749, 750
5	0.72	669, 695, 719, 749, 750
4	0.70	669, 695, 719, 749
3	0.68	669, 719, 749
2	0.54	719, 749
1	0.01	749

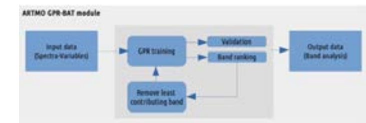
Same trend as observed as for *R*:

- ✓ **suboptimal when using many bands.**
- ✓ **Best results with 3-20 bands.**
- ✓ **3 bands at least needed to reach stable results.**
- ✓ **2-bands poor results**

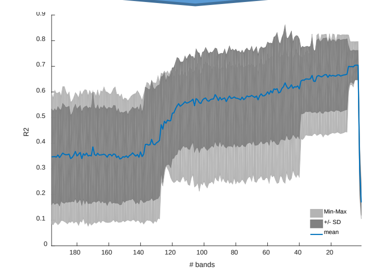
Conclusions

GPR-BAT introduced in ARTMO's MLRA toolbox.

- Iterative removal of least contributing band in GPR model development



- **GPR-BAT automatically delivers most sensitive bands of any spectral + variable dataset.**



- Various hyperspectral datasets analyzed. **GPR-BAT** results suggest:

  ✓ ***Using all bands never best result***

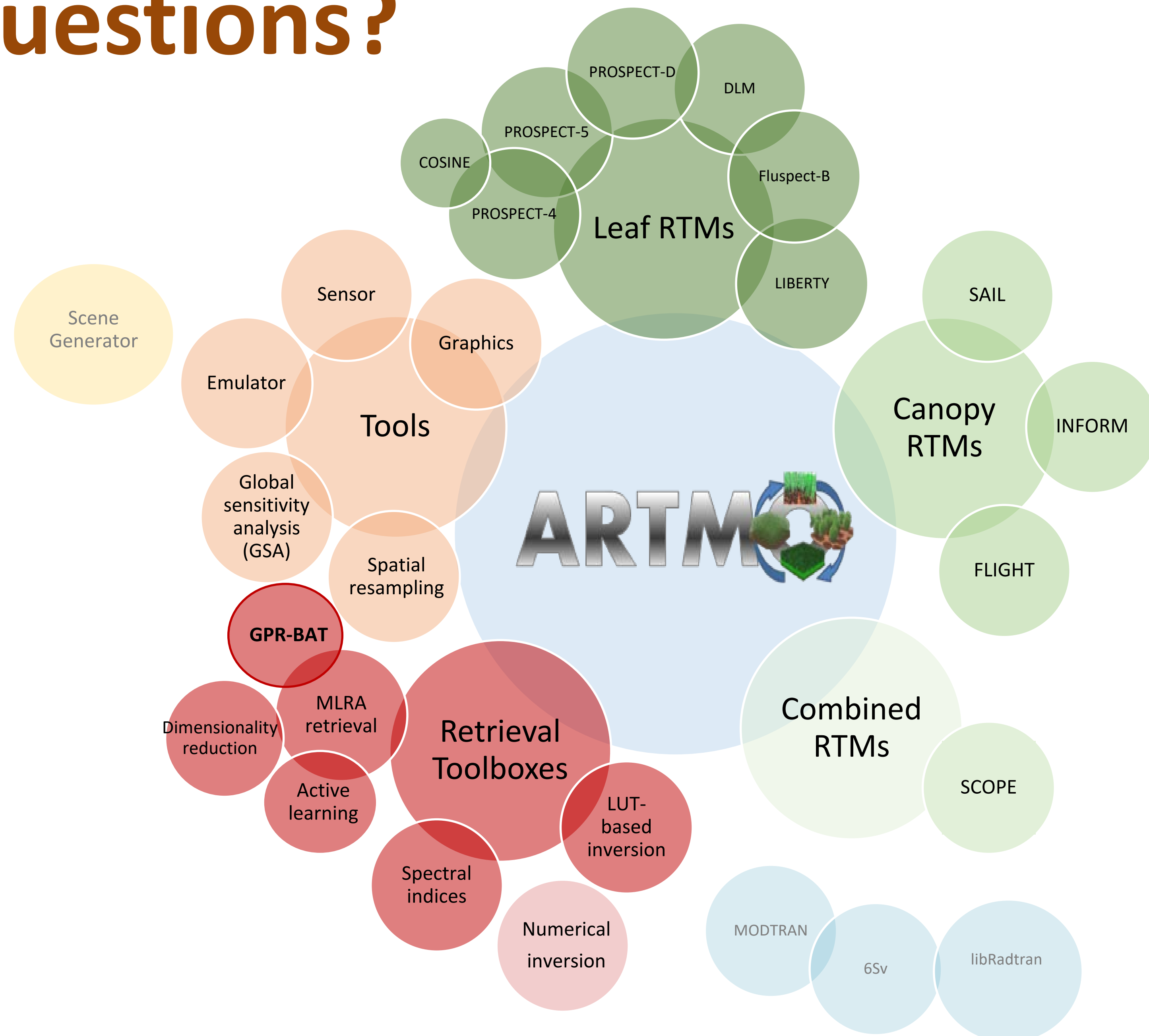
  ✓ ***Worst is using 1 band, but also 2 bands (<u>vegetation indices</u>) suboptimal.***

  ✓ ***Optimized prediction with 4-9 bands.***

  ✓ ***In the MLRA toolbox, the best performing GPR model can be applied to an image (map + uncertainty map)***

Verrelst J., Rivera, J.P., Gitelson, A., Delegido, J., Moreno, J., Camps-Valls, G., (2016). [Spectral band selection for vegetation properties retrieval using Gaussian processes regression](). International Journal of Applied Earth Observation and Geoinformation, 52, p. 554-567.

Questions?



Thanks

http://ipl.uv.es/artmo/