OTK Beta Program – Data Analysis

by Yagiz Sutcu

January 29th, 2016

Similarity Metrics

- Traditional methods for reference-library searching are typically based on the assessment of similarity metrics calculated via peak table comparisons, or more commonly, from those generated by full spectrum comparisons.
- Full spectrum approaches typically generate a “hit quality index” (HQI) between the unknown spectrum and each library spectrum. The HQI can be calculated based on Euclidean distance, median absolute deviation, or perhaps, the correlation coefficient between the test spectrum and each library spectrum.
- While the Euclidean Distance-based algorithms are generally a good first choice for spectral searching, the First Derivative-based algorithms are preferred when peak-template comparisons are considered.

Normalization

- There are two commonly used methods to normalize spectral data:
  - *Dot product normalization:* which essentially normalizes the spectrum based on the total area under the curve (mostly preferred when using full spectra comparisons)
  - *Scaling normalization:* which normalizes the spectrum based on the height of a peak of interest (preferred when using peak-template matching-based spectra comparisons)
- In this study, AUC equalization-based normalization for wavelength range [420-710 nm] is selected
- Excitation peak (laser peak around 405nm) height equalization-based normalization seems to be problematic due to some data quality issues

Statistical Analysis

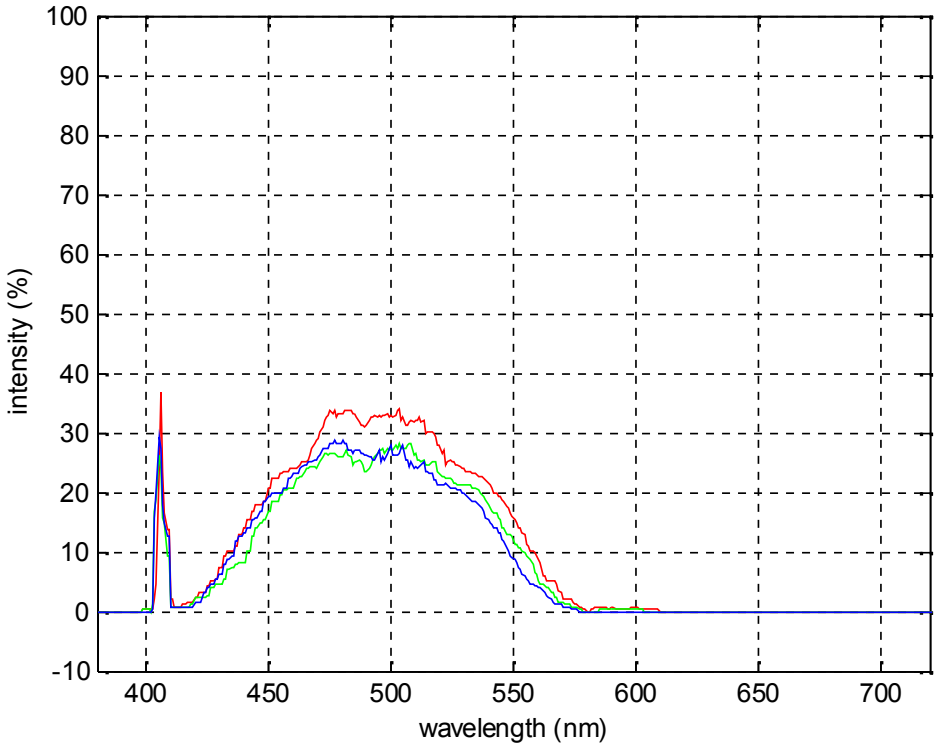
- So far 8 user have uploaded 3 spectra from 5 different samples (3 different engine oils, diesel and crude)
- Statistical Analysis:
  - *Comparing same kind of samples*
    - *Intra-user and Inter-user*
  - *Comparing different kind of samples*
    - *Intra-user and Inter-user*

Comparing Same Kind of Samples

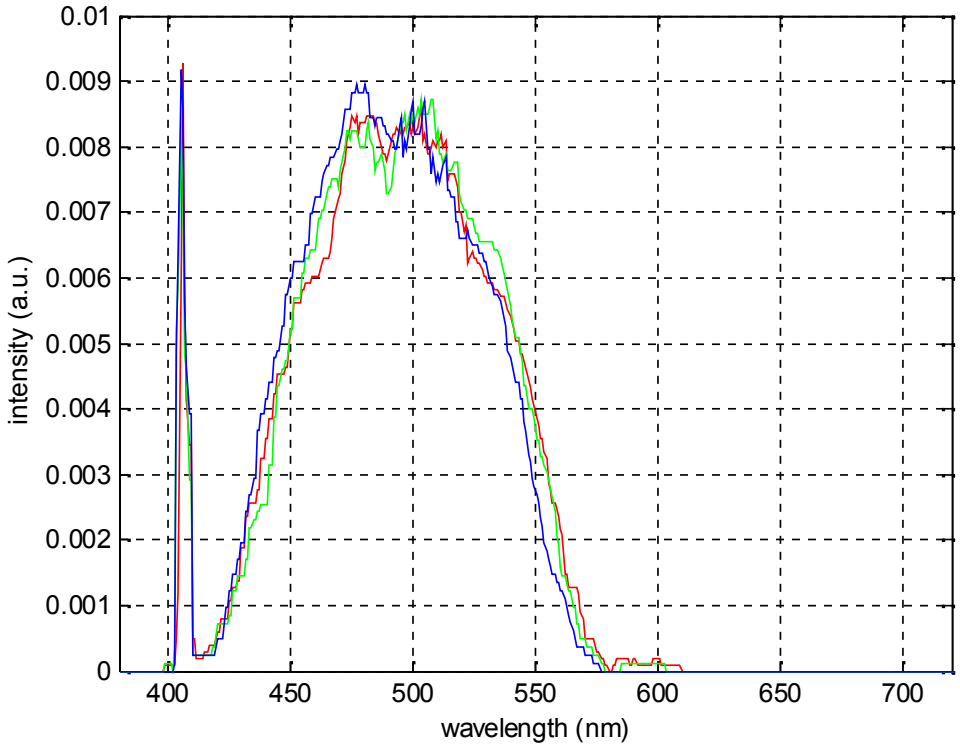
- *Intra-user Analysis:* 3 comparisons from each user give 24 error values for the same type of sample
- *Inter-user Analysis:* 8 spectra (1 spectrum from each user) give 28 error values for the same type of sample
- Mean, variance, kurtosis and skewness values of distance values are compared before and after normalization

5W30 Engine Oil

3 samples - UserID 1



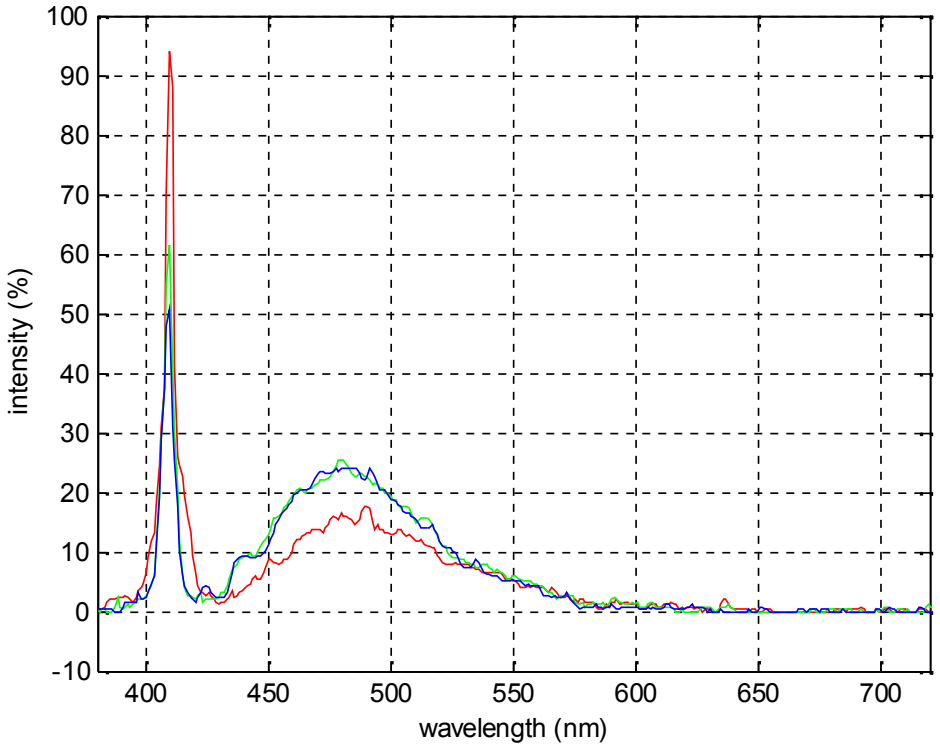
NOT normalized



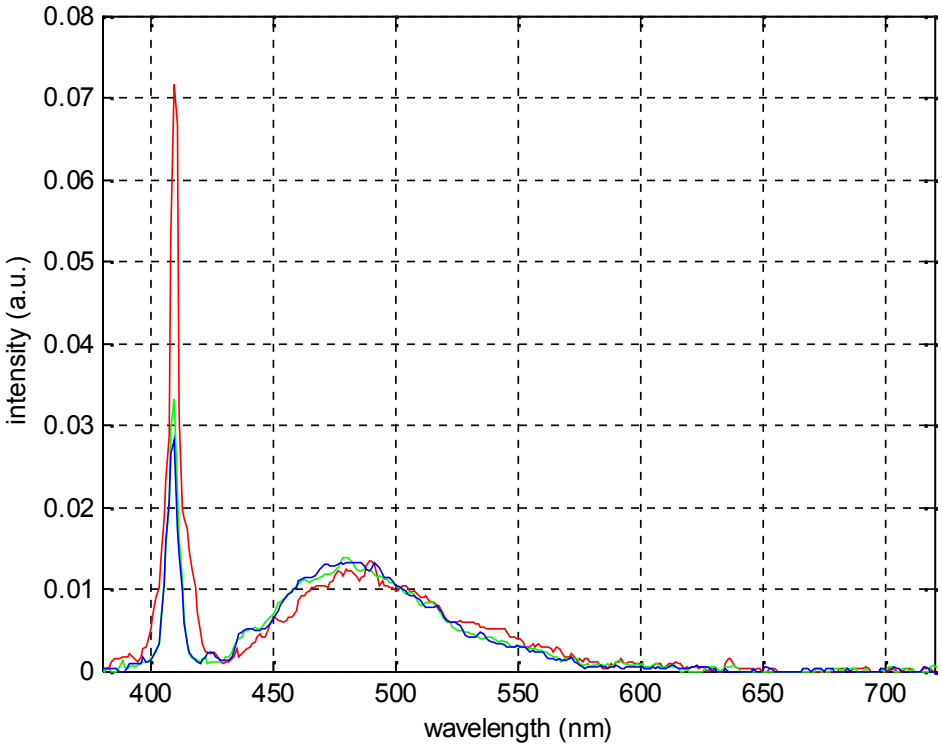
Normalized by equalizing AUC
wavelength range [420-710 nm]

5W30 Engine Oil

3 samples - UserID 2



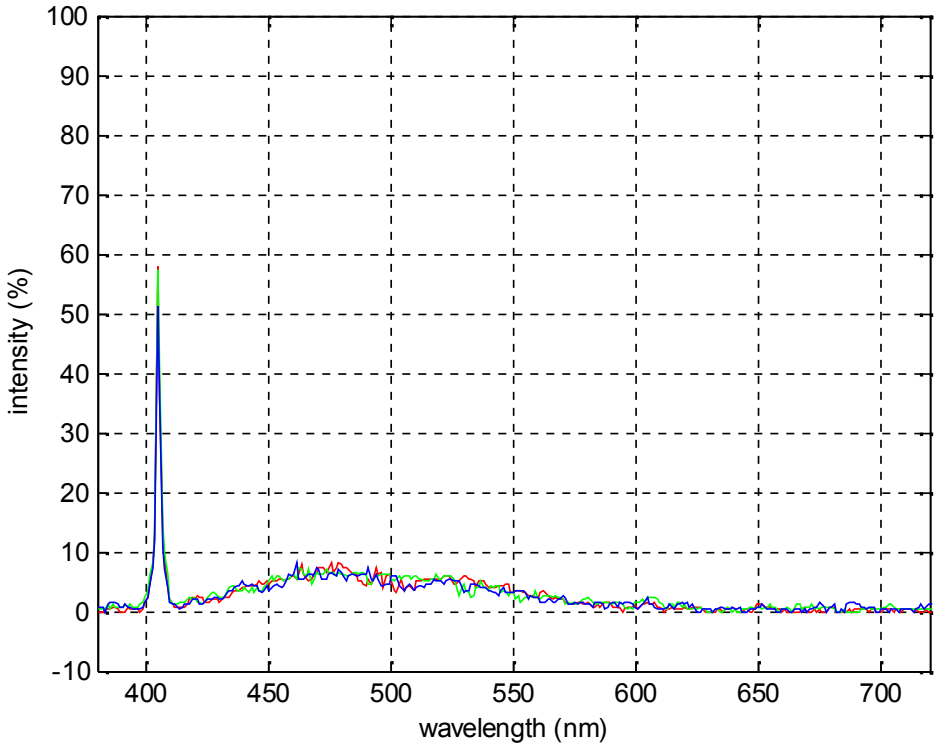
NOT normalized



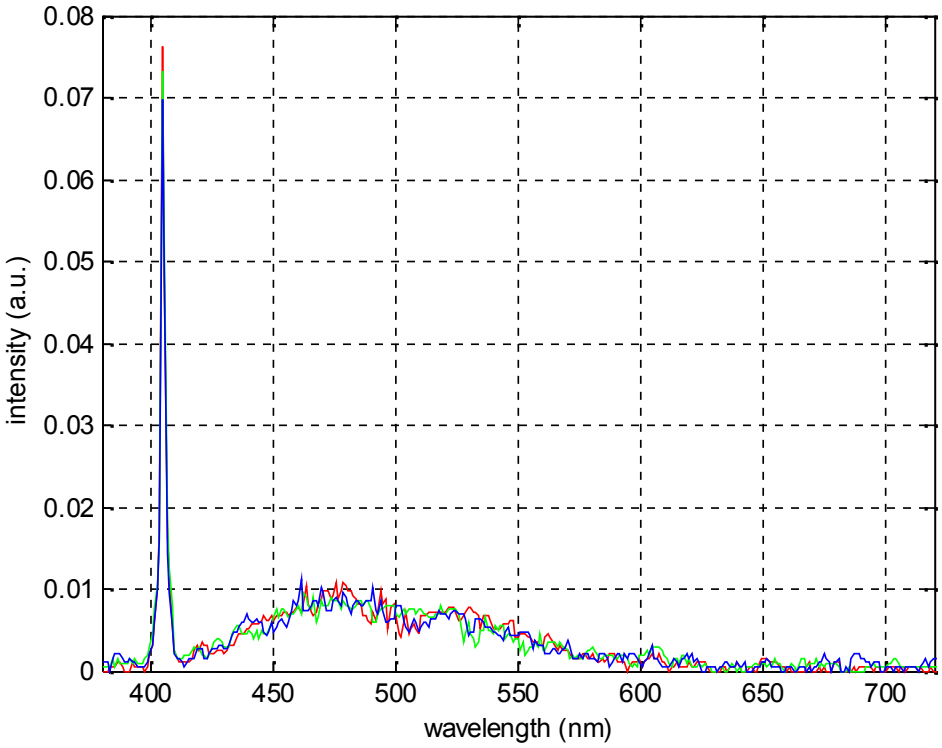
Normalized by equalizing AUC
wavelength range [420-710 nm]

5W30 Engine Oil

3 samples - UserID 3



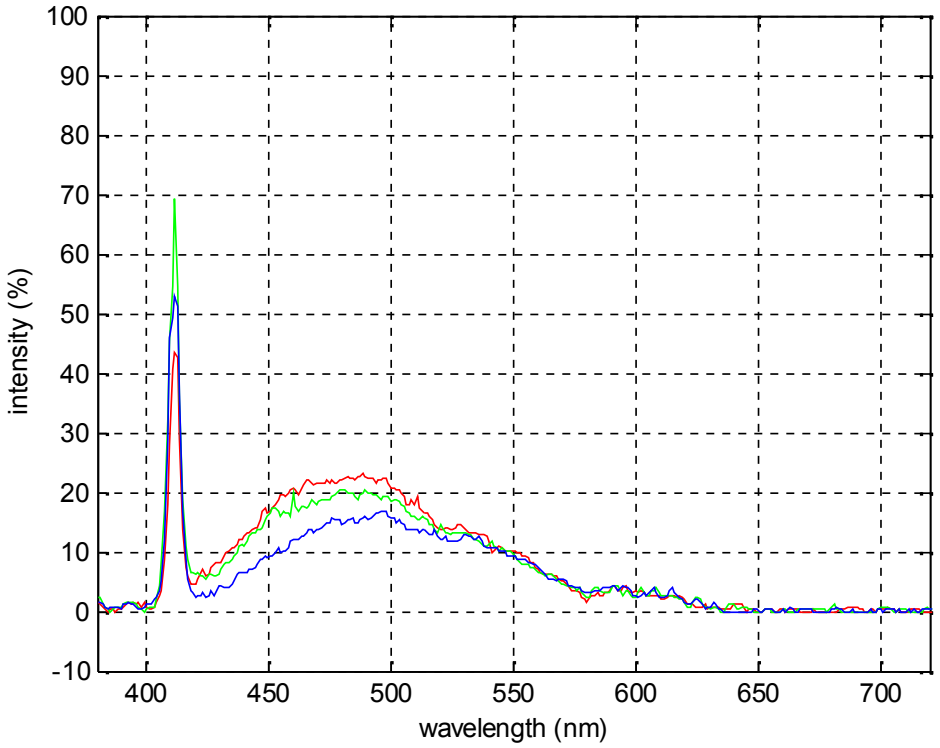
NOT normalized



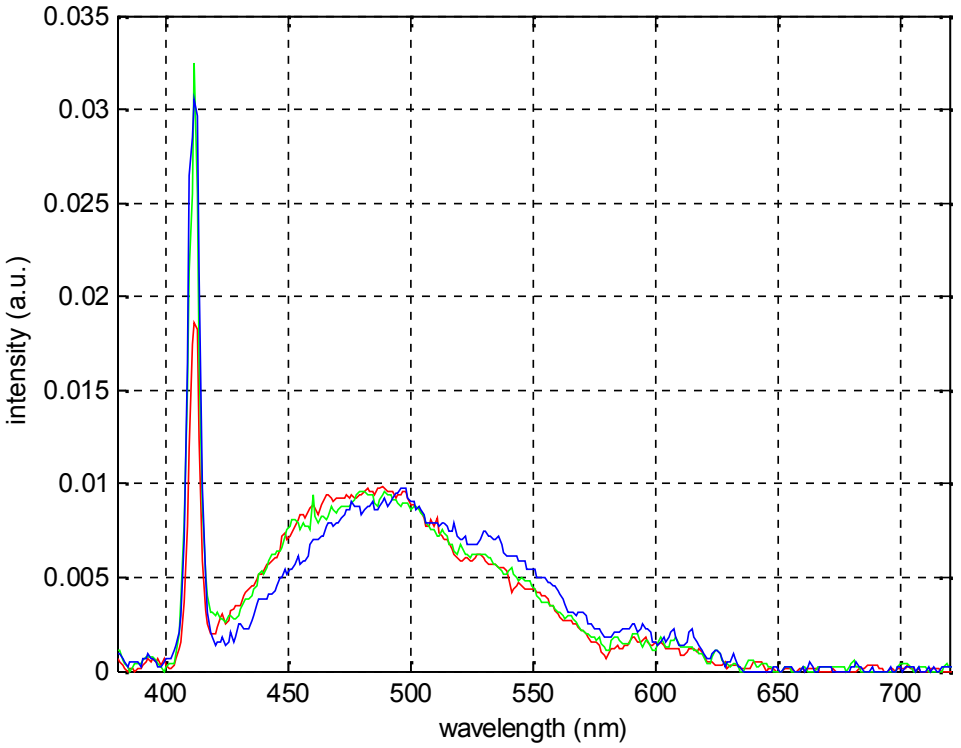
Normalized by equalizing AUC
wavelength range [420-710 nm]

5W30 Engine Oil

3 samples - UserID 4



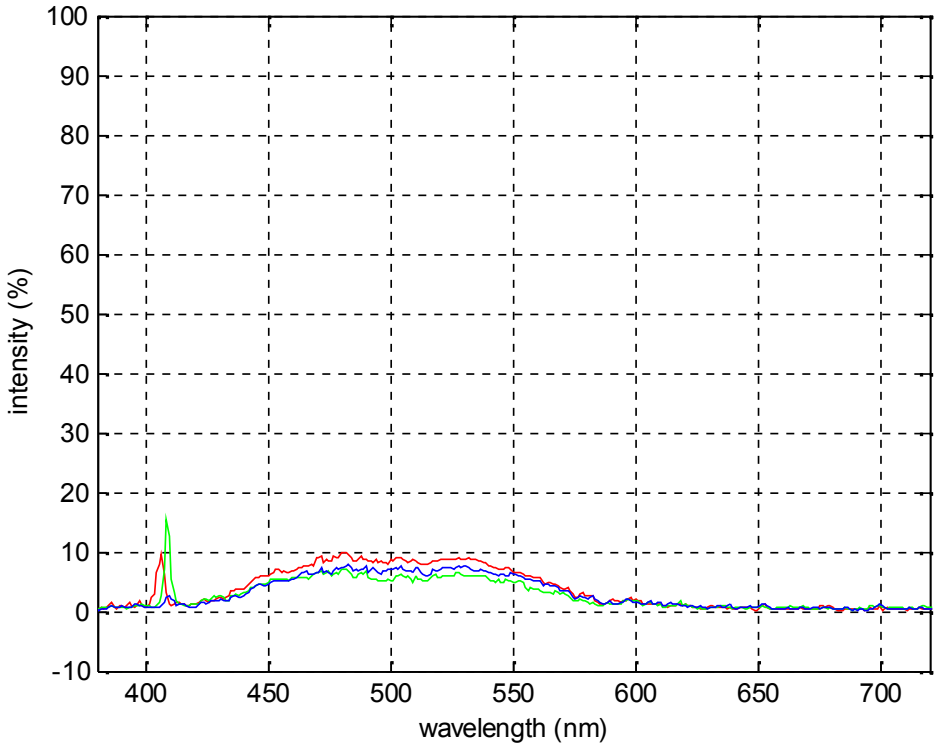
NOT normalized



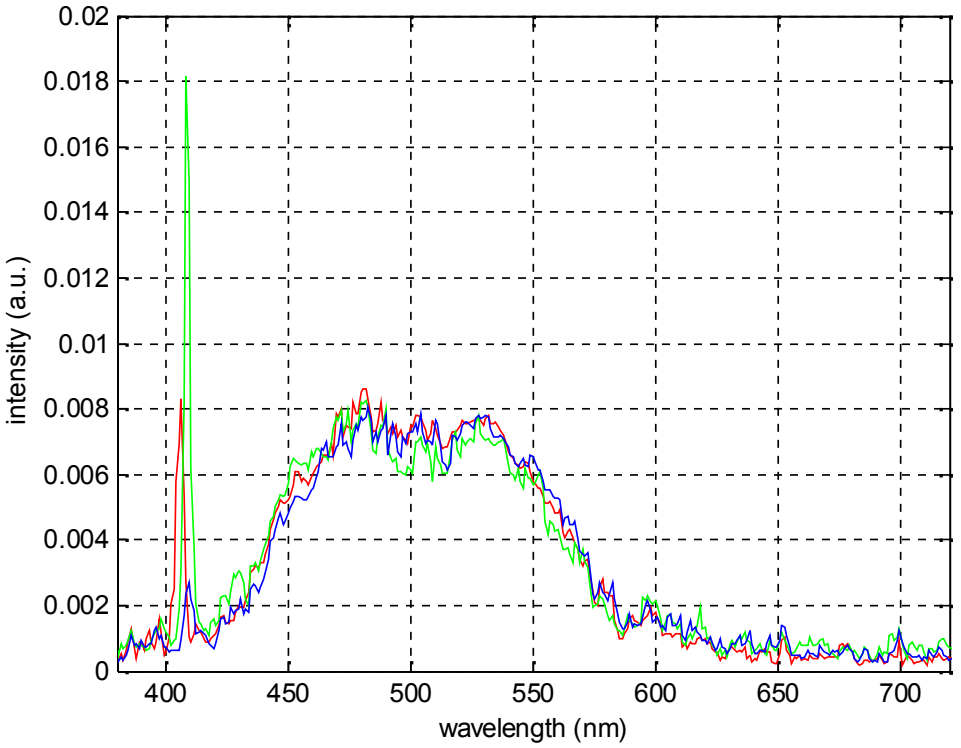
Normalized by equalizing AUC
wavelength range [420-710 nm]

5W30 Engine Oil

3 samples - UserID 5



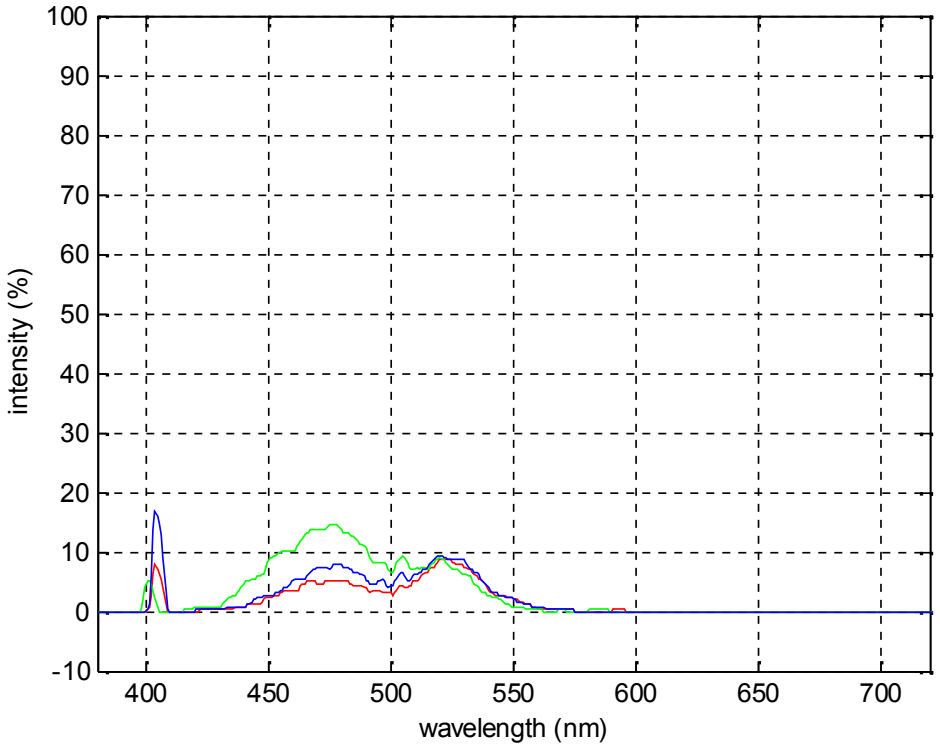
NOT normalized



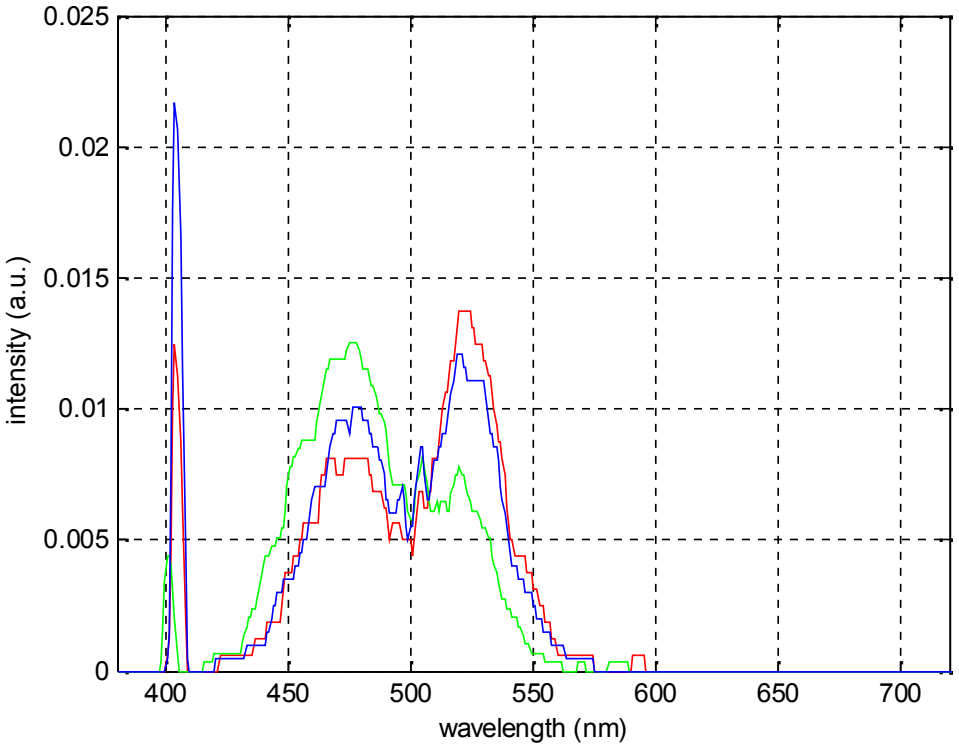
Normalized by equalizing AUC
wavelength range [420-710 nm]

5W30 Engine Oil

3 samples - UserID 6



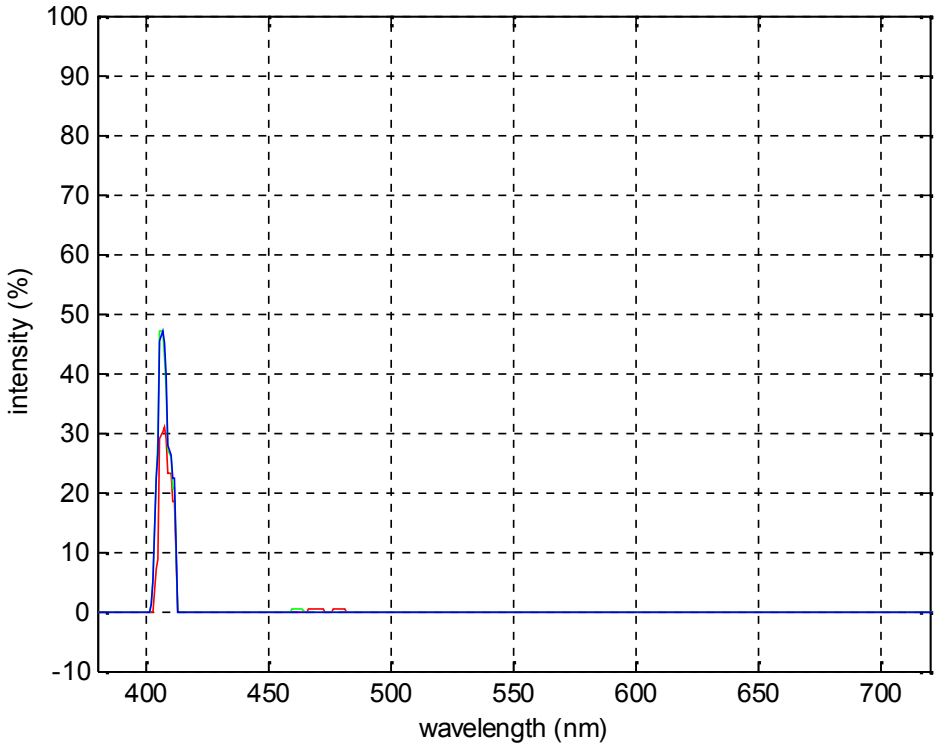
NOT normalized



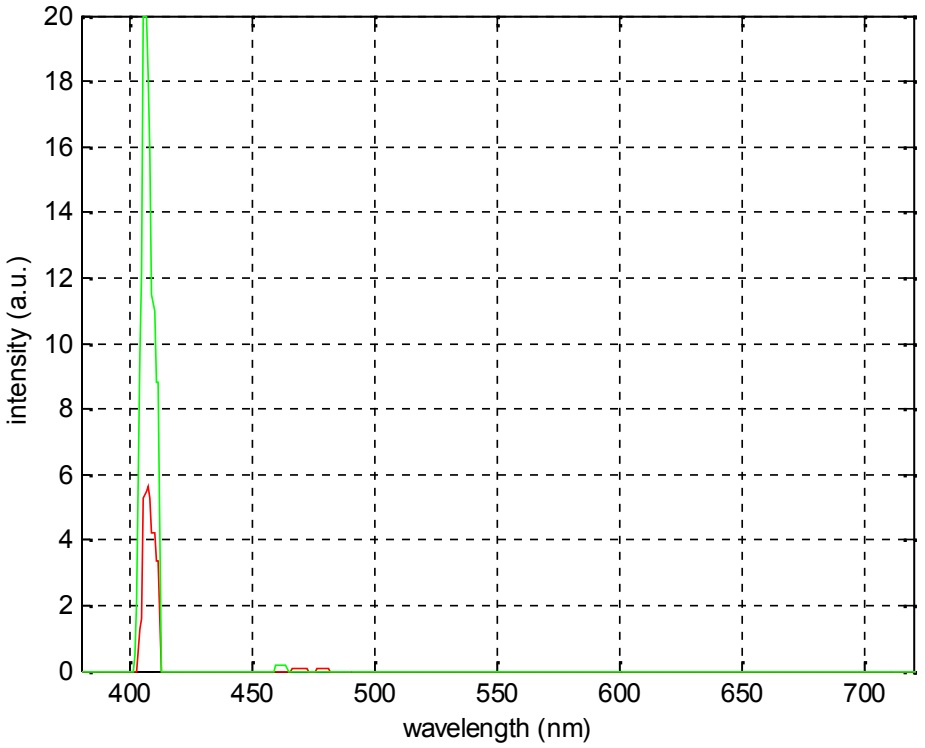
Normalized by equalizing AUC
wavelength range [420-710 nm]

5W30 Engine Oil

3 samples - UserID 7*



NOT normalized

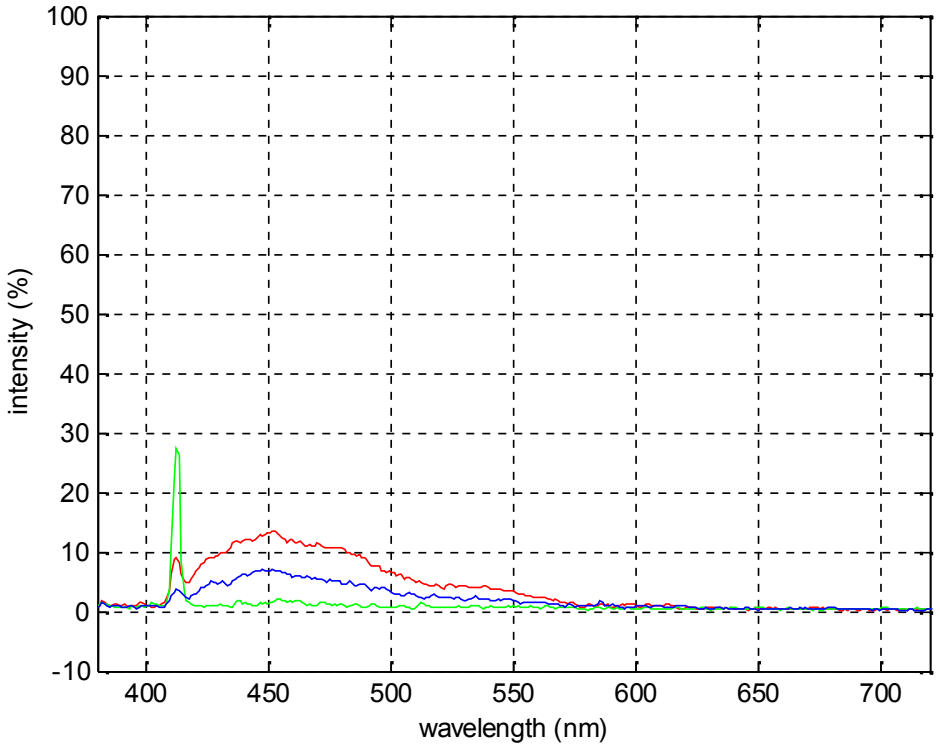


Normalized by equalizing AUC
wavelength range [420-710 nm]

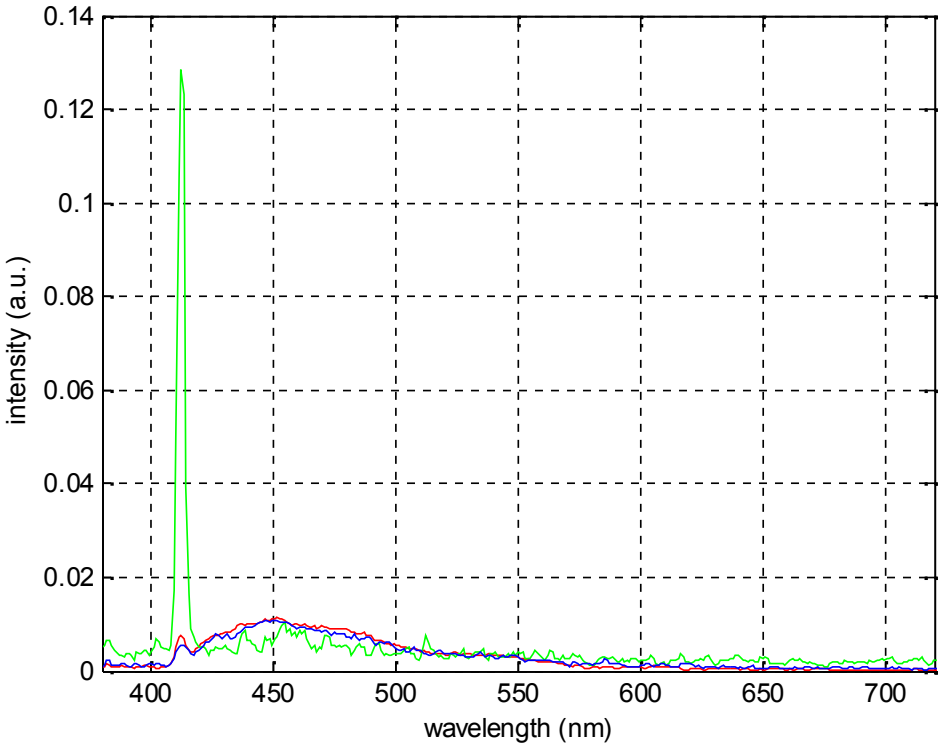
* This user's data has severe quality issues and not used in error analysis

5W30 Engine Oil

3 samples - UserID 8



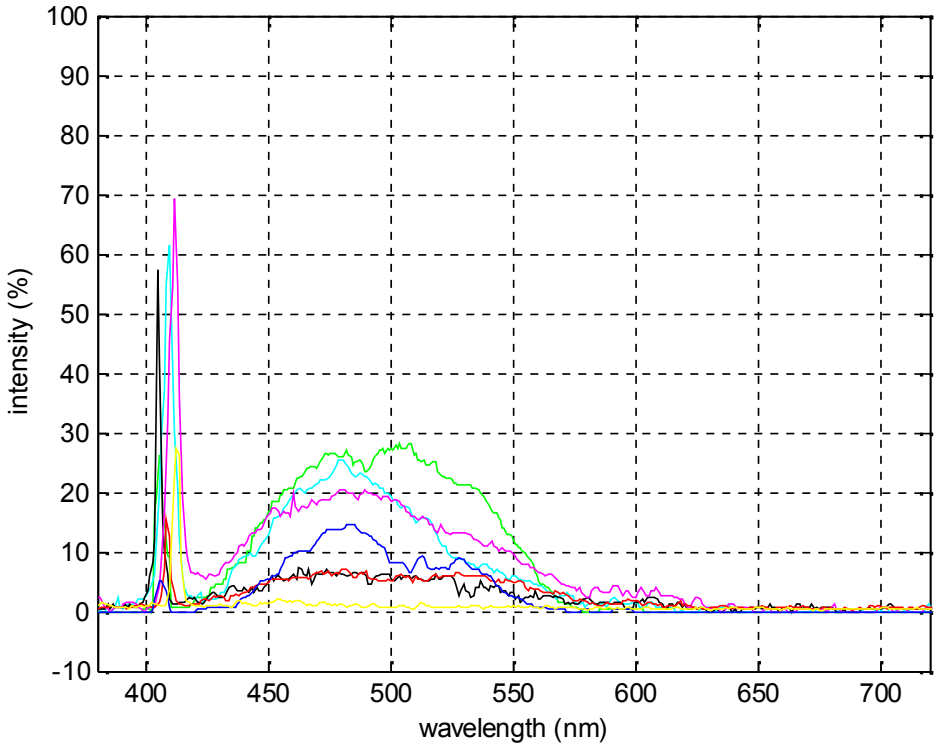
NOT normalized



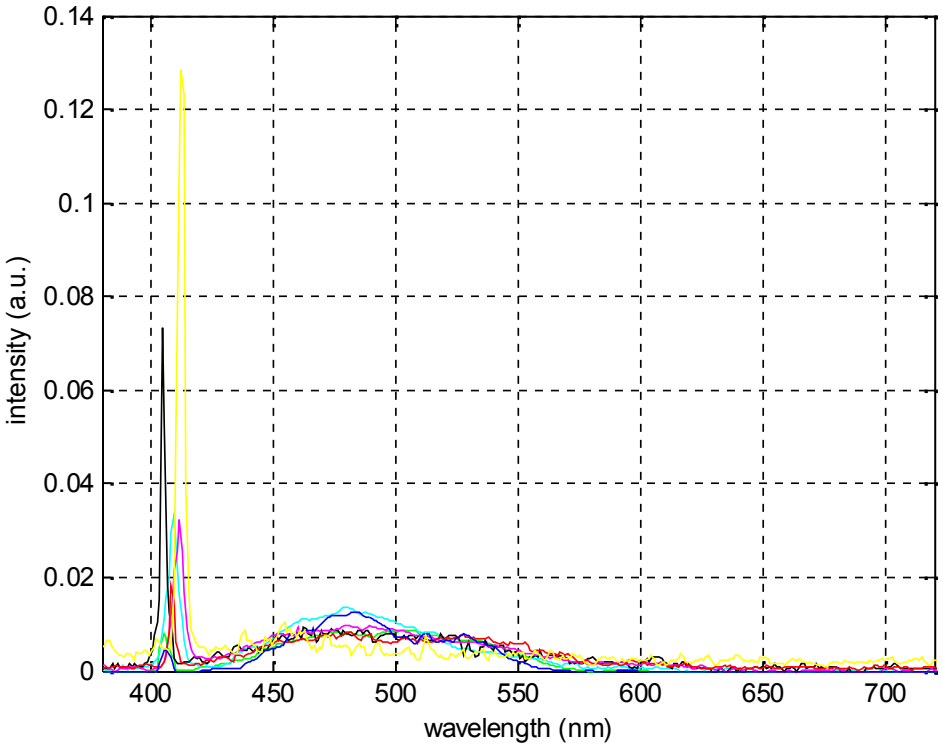
Normalized by equalizing AUC
wavelength range [420-710 nm]

5W30 Engine Oil

7 samples from all users*



NOT normalized



Normalized by equalizing AUC
wavelength range [420-710 nm]

* User 7's data has severe quality issues and not used in error analysis

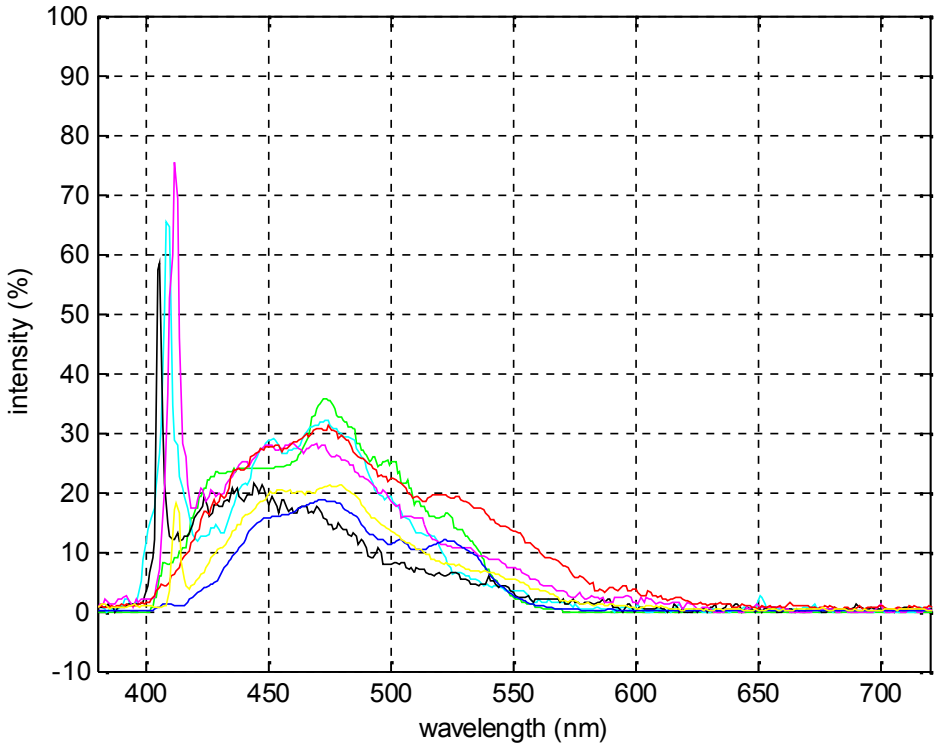
Error Analysis - 5W30 Engine Oil

	Intra-User		Inter-User	
	before norm.	after norm.	before norm.	after norm.
mean	99.3660	0.0170	374.4279	0.0607
std	58.0702	0.0115	221.6321	0.0271
kurtosis	1.9581	3.4754	1.8151	2.1378
skewness	0.3502	1.2256	0.3300	0.3033

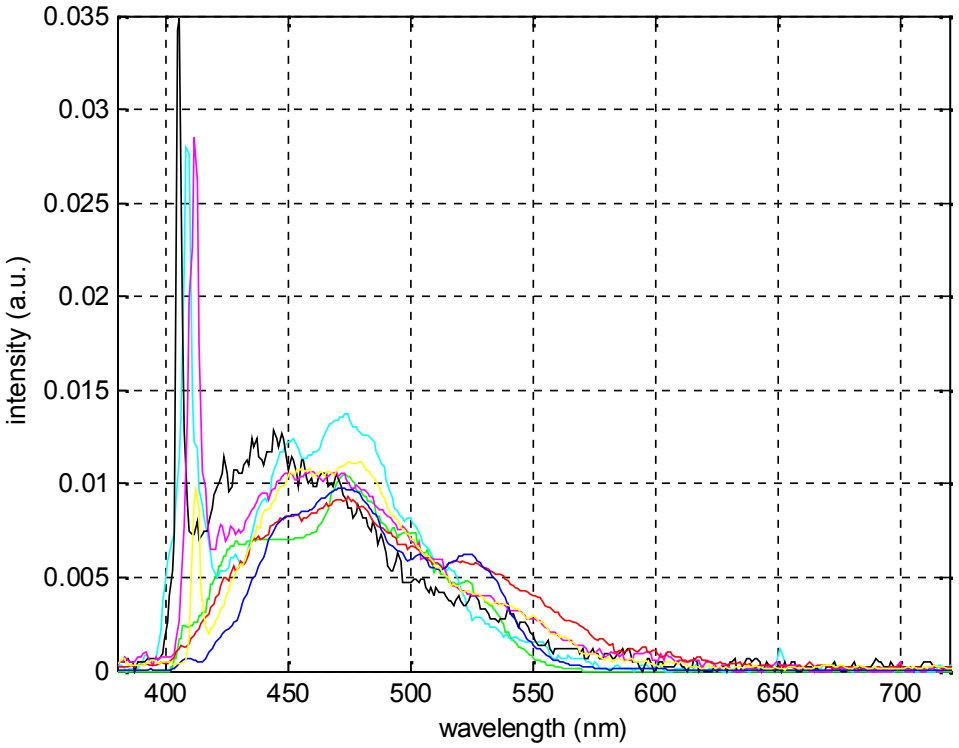
- As expected, mean and standard deviation values of error are much higher for inter-user error distribution. Although this is expected, it's statistically significantly large.
- Some of the contributing factors for the intra vs. inter user performance degradations include but are not limited to:
  - Calibration errors between users
  - Hardware/design differences
  - Instability of spectrometer, OTK attachment and connection between the two
  - Ambient conditions
  - …
- While the effect of normalization in terms of kurtosis evaluation can be considered positive (i.e. error distribution becomes more similar to normal distribution after normalization), skewness-wise there is some degeneration for intra –user case and not significant change for inter-user case.

20W50 Engine Oil

7 samples from all users*



NOT normalized

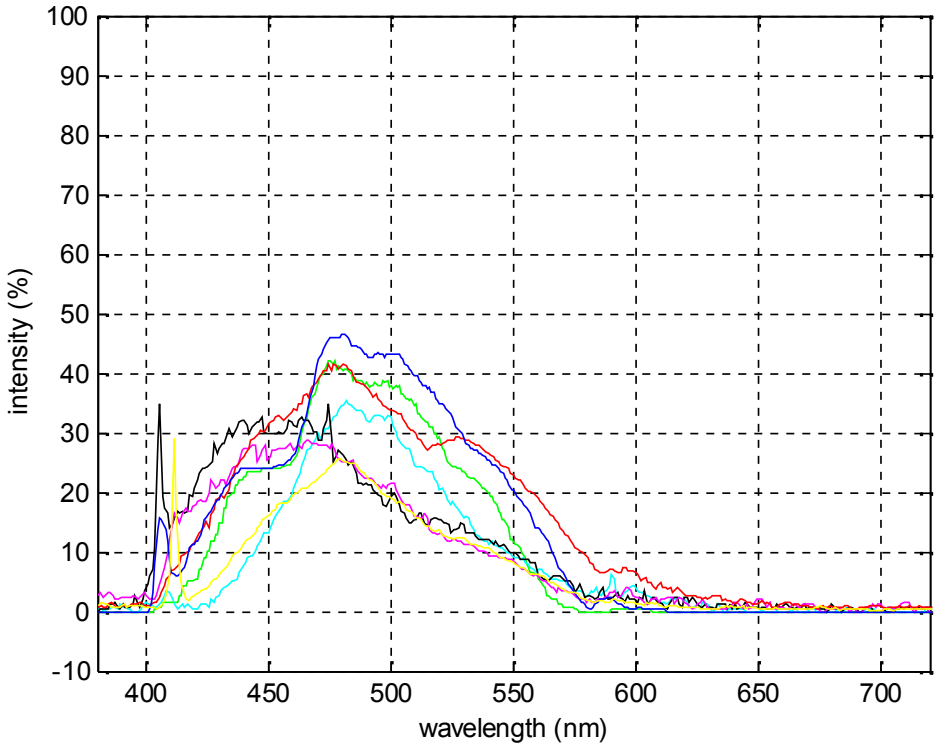


Normalized by equalizing AUC
wavelength range [420-710 nm]

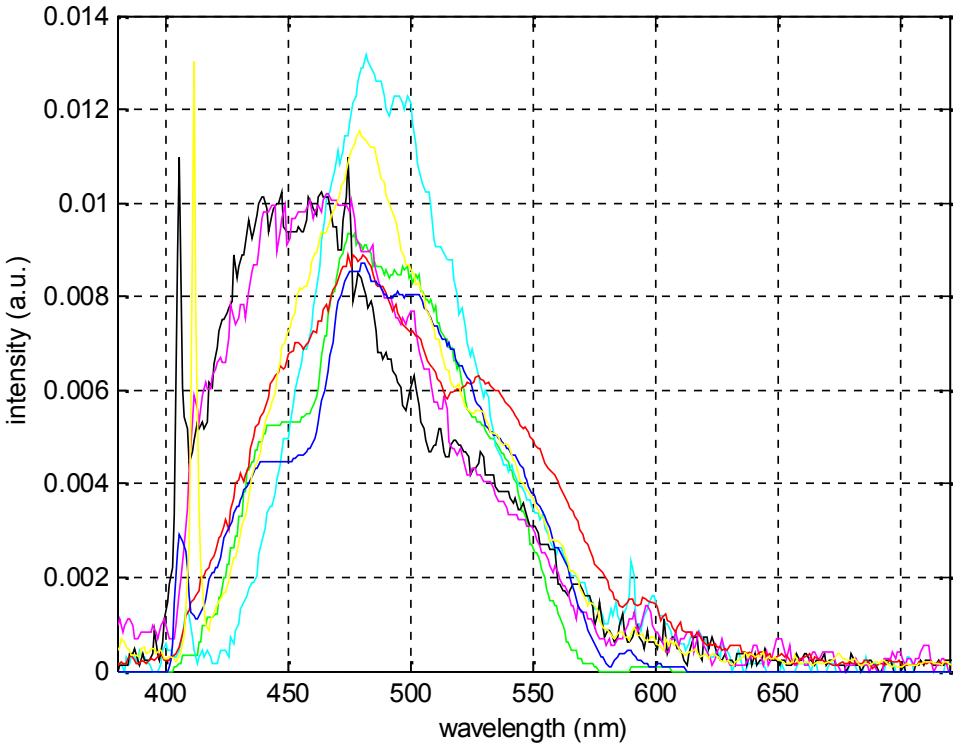
* User 7's data has severe quality issues and not used in error analysis

80W90 Engine Oil

7 samples from all users*



NOT normalized

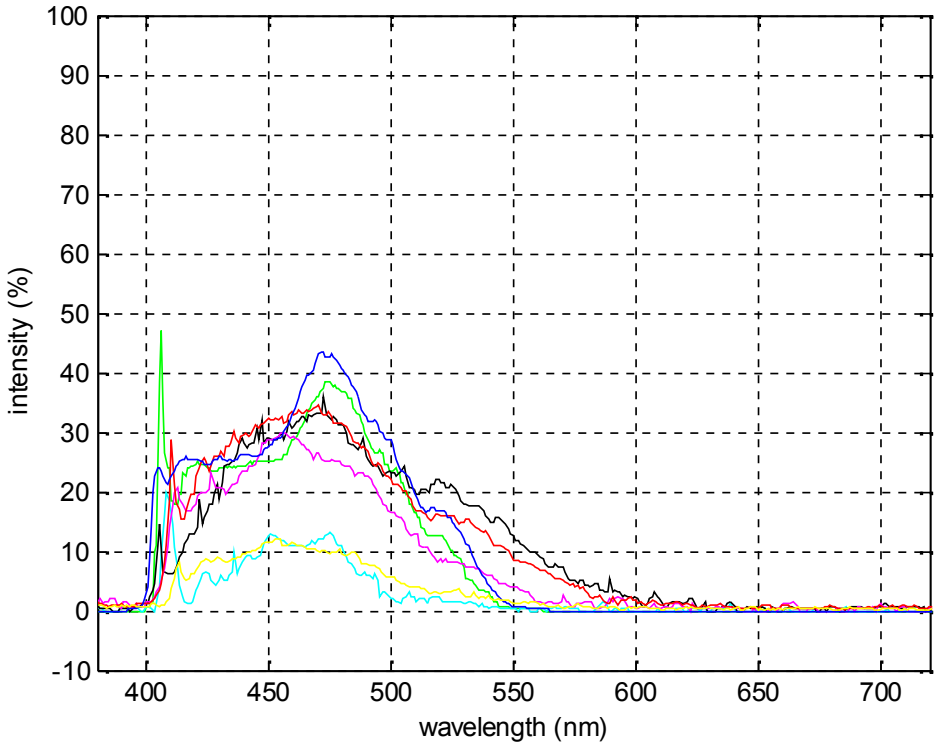


Normalized by equalizing AUC
wavelength range [420-710 nm]

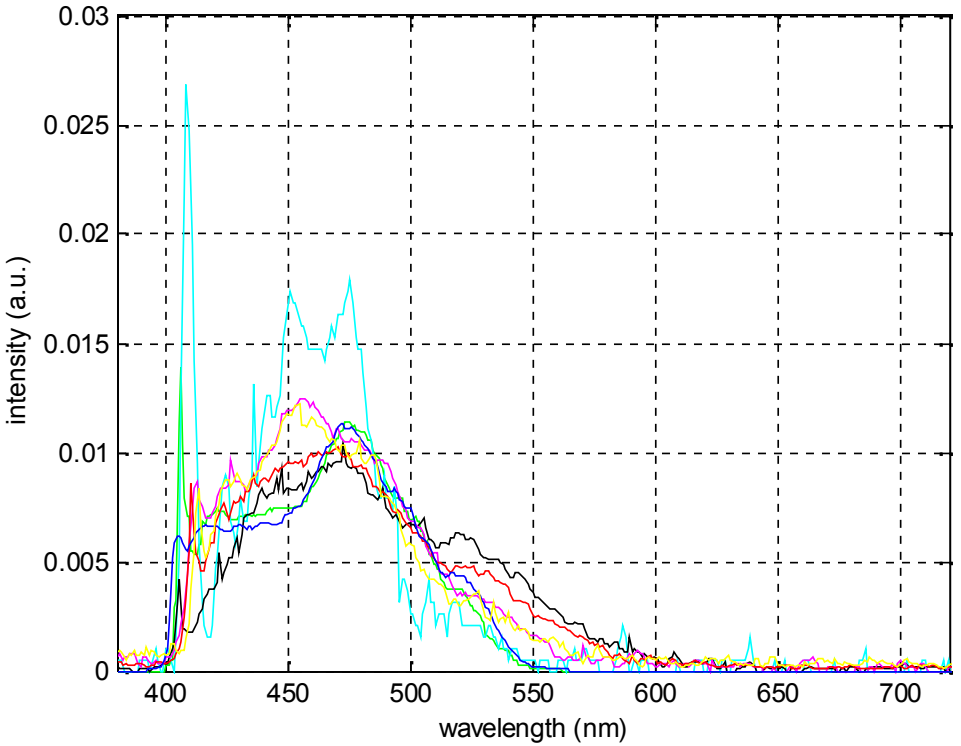
* User 7's data has severe quality issues and not used in error analysis

Diesel

7 samples from all users*



NOT normalized

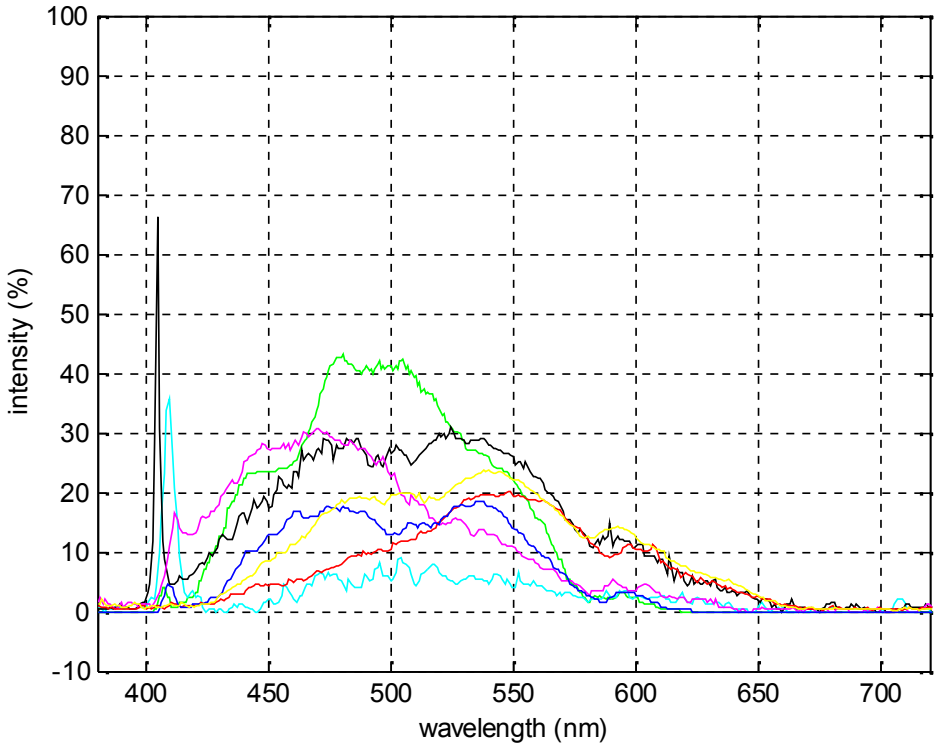


Normalized by equalizing AUC
wavelength range [420-710 nm]

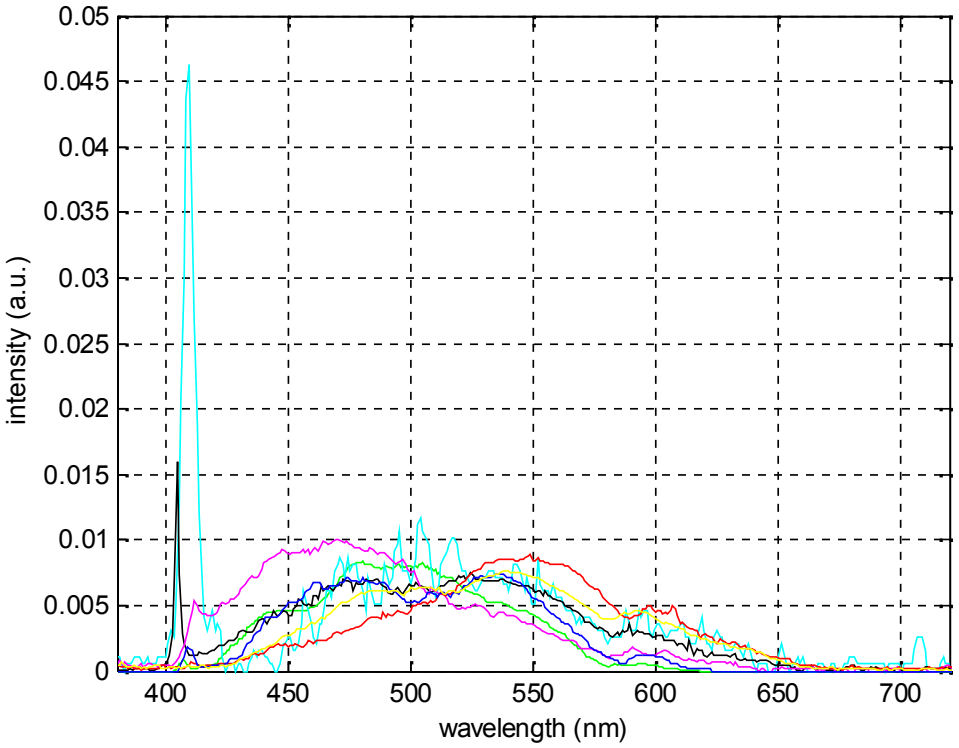
* User 7's data has severe quality issues and not used in error analysis

Crude Oil

7 samples from all users*



NOT normalized



Normalized by equalizing AUC
wavelength range [420-710 nm]

* User 7's data has severe quality issues and not used in error analysis

Error Analysis – All Types

	INTRA-USER									
	5W30		20W50		80W90		DIESEL		CRUDE	
	before norm.	after norm.	before norm.	after norm.	before norm.	after norm.	before norm.	after norm.	before norm.	after norm.
mean	99.3660	0.0170	130.7591	0.0121	164.8759	0.0109	77.6354	0.0074	158.2128	0.0119
std	58.0702	0.0115	97.3449	0.0063	161.1741	0.0086	50.6212	0.0060	146.0595	0.0125
kurtosis	1.9581	3.4754	6.7127	3.1793	6.4884	3.4734	3.2530	4.3543	2.1782	4.4752
skewness	0.3502	1.2256	1.8551	0.8343	2.1238	1.2166	1.0569	1.5662	0.8060	1.5893
	INTER-USER									
	5W30		20W50		80W90		DIESEL		CRUDE	
	before norm.	after norm.	before norm.	after norm.	before norm.	after norm.	before norm.	after norm.	before norm.	after norm.
mean	374.4279	0.0607	480.6473	0.0633	646.8477	0.0631	600.7056	0.0715	553.8771	0.0518
std	221.6321	0.0271	196.9897	0.0283	320.9634	0.0271	227.1975	0.0292	232.1696	0.0266
kurtosis	1.8151	2.1378	1.9490	2.0034	1.7829	1.8611	2.5031	2.3782	2.1455	1.8140
skewness	0.3300	0.3033	0.3850	0.5267	0.3119	0.2573	-0.2002	0.2849	-0.0503	0.2721
	Increase of the error distribution parameters (in %) from INTRA to INTER-USER									
mean	276.82	257.06	267.58	423.14	292.32	478.90	673.75	866.22	250.08	335.29
std	281.66	135.65	102.36	349.21	99.14	215.12	348.82	386.67	58.96	112.80

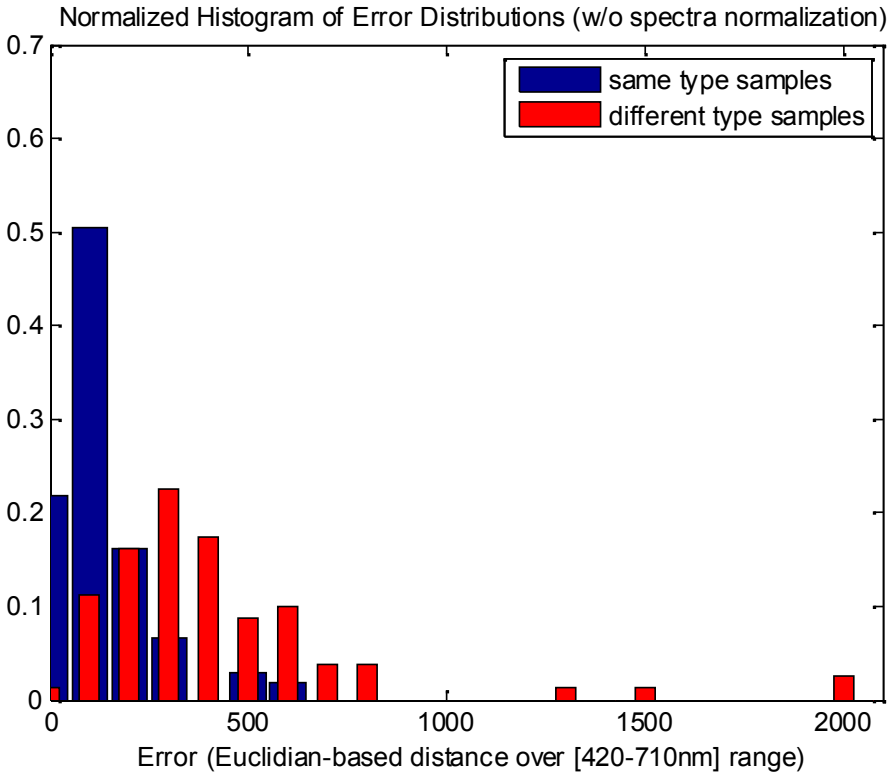
- Similar (to 5W30 engine oil analysis) observations/conclusions valid for all 5 kind of samples
- In terms of the increase in error statistics (in terms of mean and std), diesel is the worst
- Normalization (in most cases) improves intra-user statistics slightly better than inter-user statistics
- …

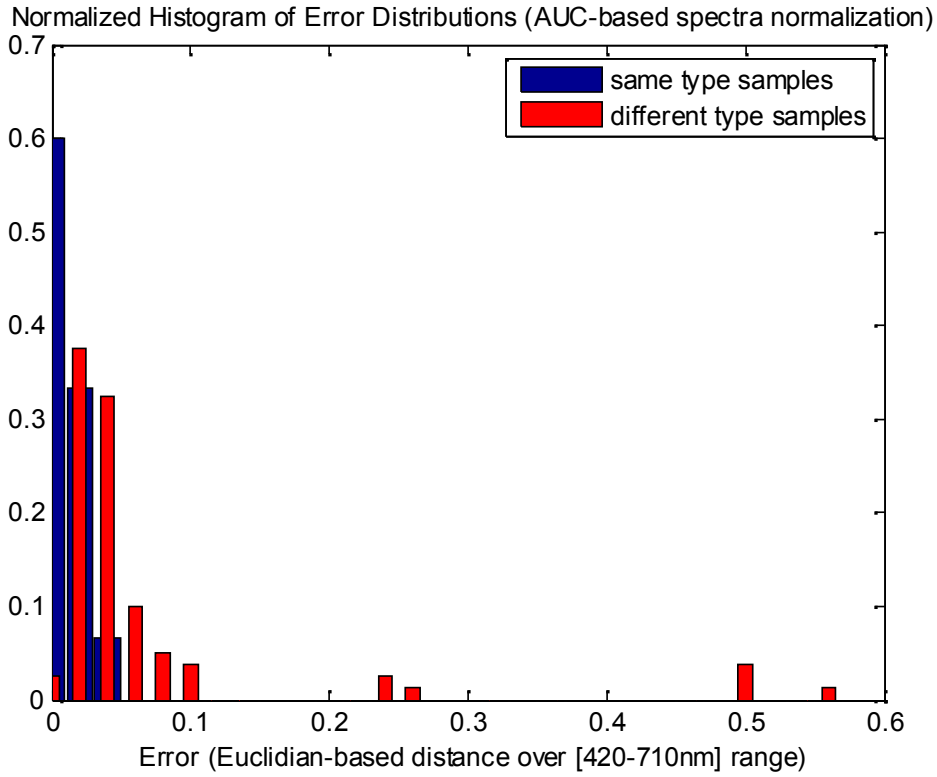
Comparing Different Kind of Samples

- *Intra-user Analysis:* 10 comparisons (5 spectra -1 spectrum from each kind of sample - from the same user) per user give total 80 error values
  - By comparing these error values with same user & same type sample error values, ***intra-user identification performance*** will be determined
- *Inter-user Analysis:* 10 comparisons (5 spectra -1 spectrum from each kind of sample - from different users) – repeated several times with randomization
  - By comparing these error values with different user & same type sample error values, ***inter-user identification performance*** will be determined

Intra-user Identification Performance

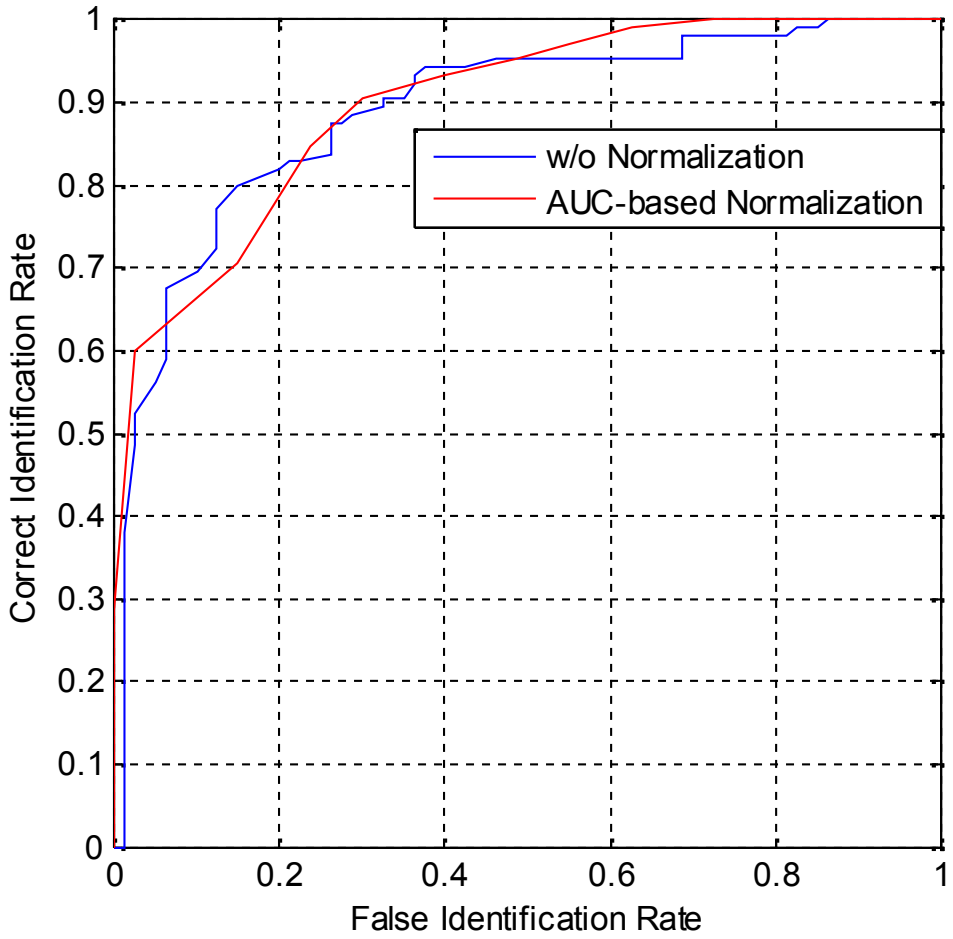
Error Histograms





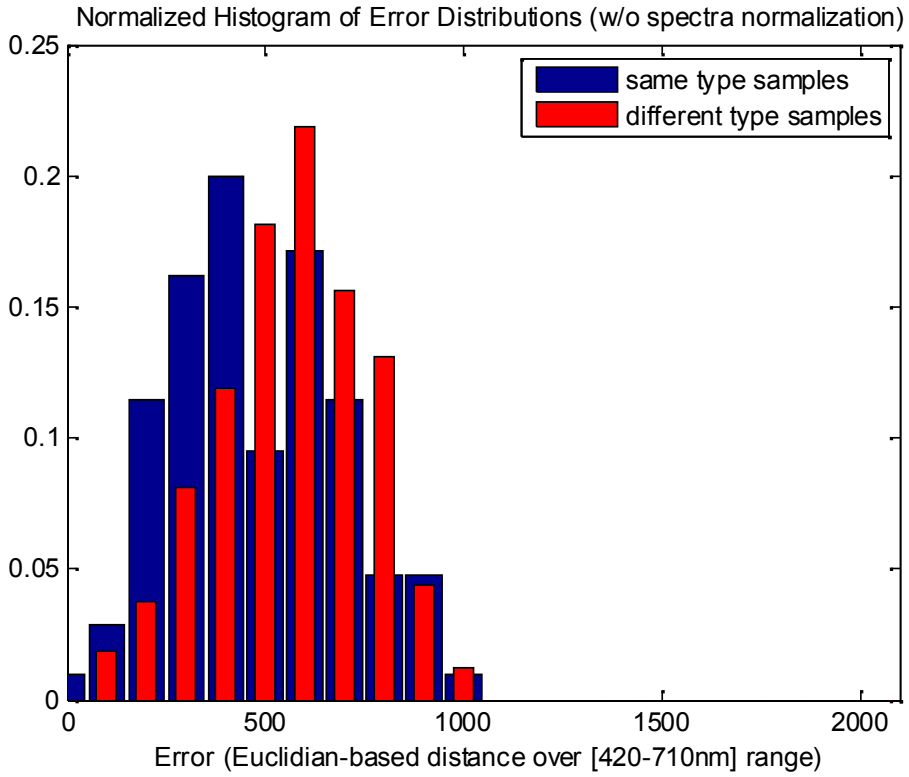
Intra-user Identification Performance

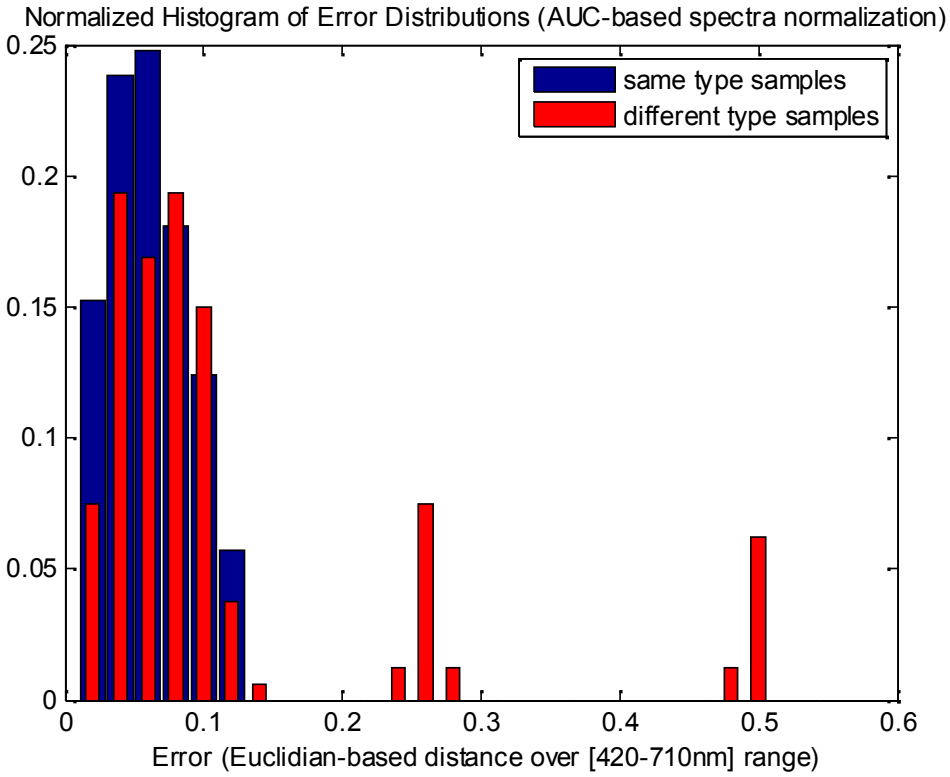
ROC Plot



Inter-user Identification Performance

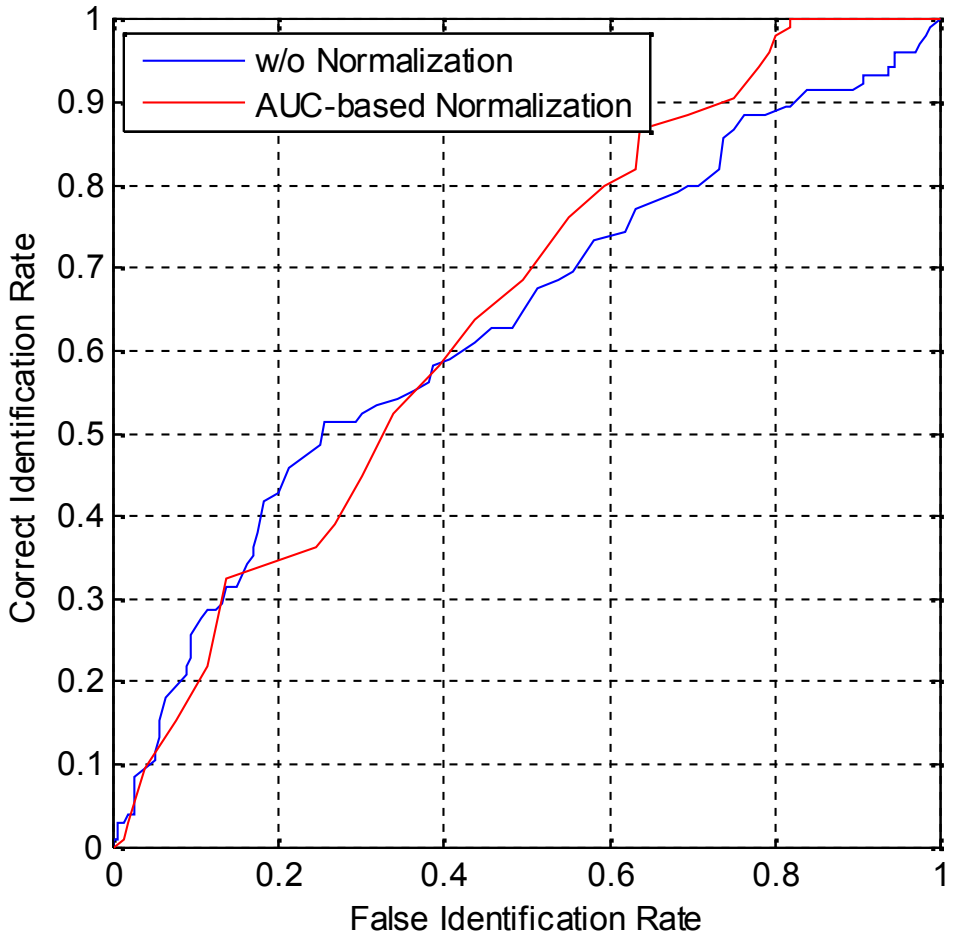
Error Histograms





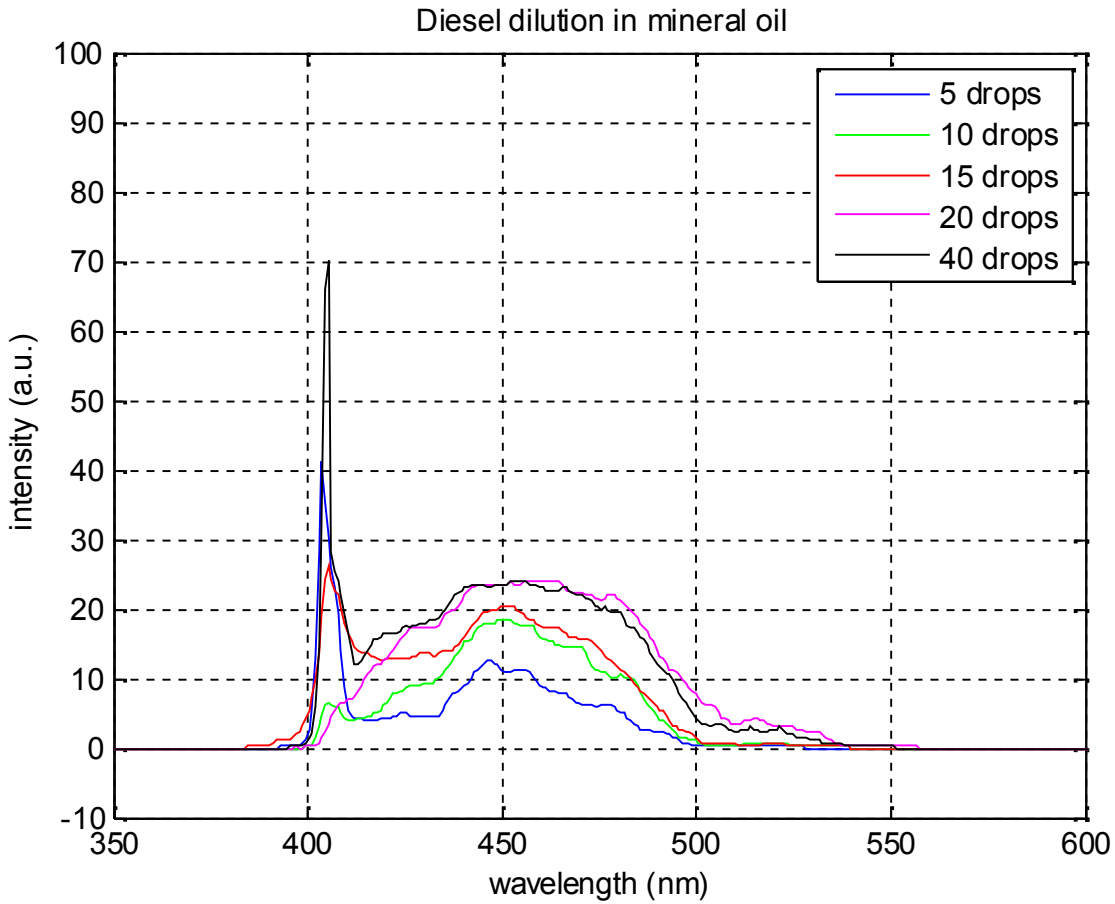
Inter-user Identification Performance

ROC Plot



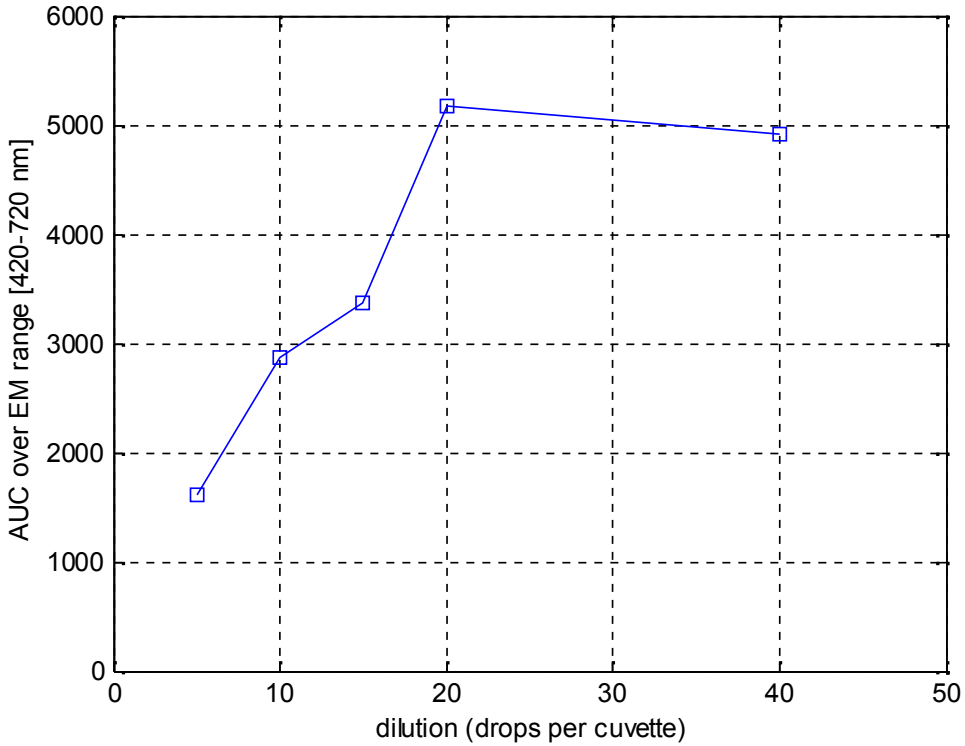
Dilution/Concentration Analysis

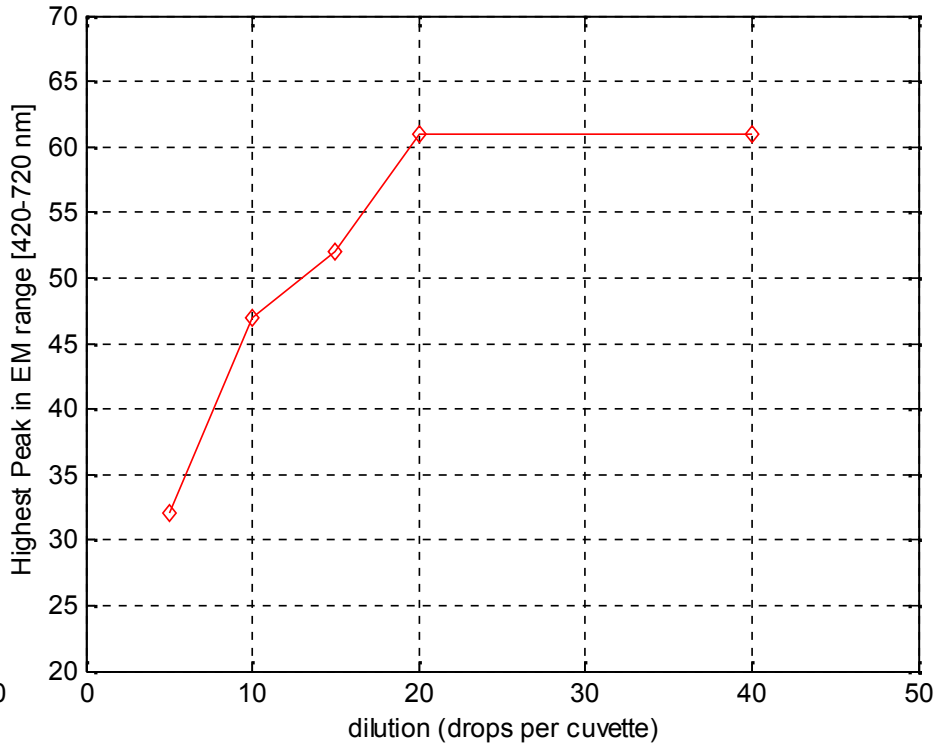
5 samples from user 6



Dilution/Concentration Analysis

5 samples from user 6





- Diesel is diluted in mineral oil and 5 different dilution ratios are tested.
- First 4 points shows almost a linear relation between concentration and AUC and peak height in EM band.
- …

Conclusions and Discussion

- Identification of these 5 different type of pollutants by comparing different users' samples is not possible.
- Same user/device-based identification performance is much better than inter-user identification. However, it cannot be considered as acceptable with current numbers. However, for this setup/design, it's not bad at all…
- Smoothing may help improve the performance in overall (needs to be tested).
- Template/cluster generation might be a good idea for database/library creation.
- …