A REAL-WORLD HYPERSPECTRAL IMAGE PROCESSING WORKFLOW FOR VEGETATION STRESS AND HYDROCARBON INDIRECT DETECTION

D. Dubucq [1]*, N. Audebert [2], V. Achard [2], A. Alakian [2], S. Fabre[2], A. Credoz [1], Ph. Deliot [2], B. Le Saux [2]

[1] TOTAL SA, 64018 PAU, France - (dominique.dubucq, anthony.credoz)@total.com
[2] ONERA, Toulouse & Saclay, France (veronique.achard, alexandre.alakian, sophie.fabre, philippe.deliot)@onera.fr

Commission III, WG III/4

KEY WORDS: hyperspectral image processing, machine learning, unmixing, hydrocarbon indirect detection

ABSTRACT:

In this work, we present the complete workflow used to acquire a large hyperspectral dataset on a western Africa historical hydrocarbon production site, and its processing. Our goal is to study how state-of-the-art hyperspectral processing techniques can help detect hydrocarbon bearing soil either of natural origin or accidental by monitoring the health of the vegetation for exploration or environmental monitoring purposes. We present our complete workflow, from acquisition, atmospheric correction, image annotation and classification using modern machine learning techniques, and show how state-of-the-art research can be applied to real-world use cases.

1. INTRODUCTION

As remote sensing technology progresses, automated data mining of Earth Observation data can be used for new applications. The objective of this work was to combine state of the art sensors and algorithms to extract knowledge from remote -

Thanks to previous studies showing that hyperspectral data can be leveraged to characterize hydrocarbon (Arellano et al., 2015), (Lassalle et al., 2018) through indices related to vegetation health (Noomen et al., 2012) and recent advances in statistical learning for hyperspectral image processing (Ben Hamida et al., 2016), it becomes feasible to investigate fully automated stressed vegetation detection from remote sensing data. However, we will see that not all the detected anomalies can be related to hydrocarbon impact.

The area of interest is located in Western Africa in a coastal region where sparse and dense vegetation with partial urbanization are present. Offshore and onshore hydrocarbons have been produced for over 50 years in that area. At the time of acquisition reclamation works were ongoing but we had the opportunity to acquire a large hyperspectral dataset before an old oil bearing well mudpit was cleaned. The objective was to develop a methodology to detect and identify surface hydrocarbons whether natural or resulting from anthropic activity, in similar environments. We present a complete hyperspectral image processing workflow using machine learning to show how state of the art remote sensing techniques can be used to extract meaningful information from such a real-world dataset.

2. DATA ACQUISITION AND PRE-PROCESSING

The hyperspectral acquisitions over area of interest were made using two HySpex cameras from NEO (Norsk Elektro Optikk): a VNIR1600 and a SWIR320m-e which have respectively 160 spectral bands in the VNIR domain [0.4 µm - 1.0 µm] and 256 bands in the SWIR domain [1.0 µm - 2.5 µm]. Airborne

hyperspectral images were acquired at an altitude of about 1750 m, which lead to a ground sampling distance of 1.3 m for the VNIR – equipped with a field of view expander – and 2.6 m for the SWIR. The field of view expander on the VNIR camera was added to cope with the high velocity of the plane (Falcon20 jet). The cameras were installed inside a pod located below the left wing, as illustrated in Fig. 1. Characteristics of these cameras are detailed in Table 1.
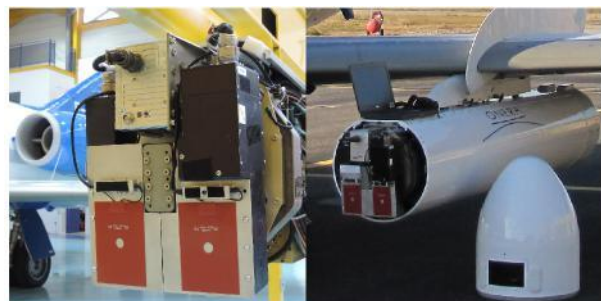

Figure 1 Camera VNIR1600 and SWIR256m-e inside the front part of the pod.

The conditions of acquisition were quite good with very few clouds. The radiometric correction was made using the NEO software which makes an offset and gain correction and produced at sensor radiance images thank to calibration parameters measured in the laboratory. Due to aerodynamic constrains of the pod, the cameras must look through a quartz window. So, another radiometric step was conducted which consists in dividing the signal by the transmission of this window as seen by each spectral band of the sensors.

Hyperspectral Radiance images were then corrected from atmospheric and environment effects using COCHISE (Miesch et al.,2005), which is based on the use of hyperspectral

---

*   Corresponding author

information combined with the radiative transfer code MODTRAN (Berk et al., 1999). The final products are ground reflectance images. The ortho-rectification step was made with PARGE software and VNIR and SWIR images were registered using GeFolki (Plyer et al., 2015). The final image is 1576 × 1000 pixels with 416 spectral bands.

Hyperspectral data analysed in the following paragraphs are extracted from a single flight line.

Main Features	Hyspex VNIR + Hyspex SWIR	
Imaging	Pushbroom	Pushbroom
Spatial pixels	1600	320
Field of view	34°	14°
Spectral range	0.4 −1 µm	1 −2.5 µm
Spectral bands	160	6 nm
Spectral sampling distance	3.7 nm	6 nm
Ground sampling distance	1.3 m	2.6 m

Table 1 Camera specifications

3. GROUND TRUTH ANNOTATION

An in-situ campaign allowed us to acquire multiple photographs, ASD field spectrometer and humidity measurements on the ground to identify the different classes and create ground truth image. The co-registered hyperspectral pixels were then manually annotated based on the ground measurements. This annotation process is long and tedious compared to cat and dog images labelling or urban environment annotation: each pixel needs to be very precisely geolocalized to match it with the ground description. Stressed vegetation has no structure or shape as a building or a road. The resulting ground truth (Fig. 2c) covers a small part of the image (42,384 pixels). Therefore, we extended the annotations based on a very strict similarity measure. A pixel with reflectance $\rho 1$ is associated to the class i if it fulfils these criteria with at least one-pixel $\rho 2$ from class i:

• Spectral reflectances have similar shapes, i.e. the spectral angle between reflectances is below a threshold:

$$E1 = \cos^{-1}((\rho 1|\rho 2)/(\|\rho 1\|*\|\rho 2\|)) < S1$$

• Reflectance amplitudes are not too different, i.e. the difference of the Euclidean norms of the reflectances is below a threshold:

$$E2 = |\|\rho 1\| − \|\rho 2\|| < S2$$

This criterion allows us to discriminate between pixels with similar shapes but different amplitudes (e.g. stressed vegetation and gray sand).

• The two spectra present the same specific absorption, if any (e.g. clay, carbonate...), i.e. the maximal difference between normalized reflectances is below a threshold:

$$E3 = \max_{\lambda} |\rho 1(\lambda)/\|\rho 1\| − \rho 2(\lambda)/\|\rho 2\|| < S3$$

While the spectral angle is a global criterion, this allows us to detect local dissimilarities between spectra.
S2 and S3 are set respectively to 0.1 and 0.05. S1 is the most critical threshold, since it defines the confidence level of the extended ground truth. We chose a conservative value of 1° for a limited but very reliable extrapolation. This clustering allows us to obtain the ground truth illustrated in Fig. 2d comprised of 157,259 annotated pixels (10% of the image).
This extended ground truth was further validated by the team that has been on the ground.

4. CLASSIFICATION

Considering that our goal is to find direct and indirect clues related to the presence of hydrocarbons, we investigate several machine learning strategies for land cover classification.

4.1 Healthy/stressed vegetation classification

Recent studies showed that vegetation health is a good indicator of soil pollution due to oil or gas (Arellano, 2015), (Noomen, 2012), (Lassalle, 2018). The differences in parenchyma structure and cell morphology due to stress induce a variable response in the near-infrared (NIR), while the variations in plant
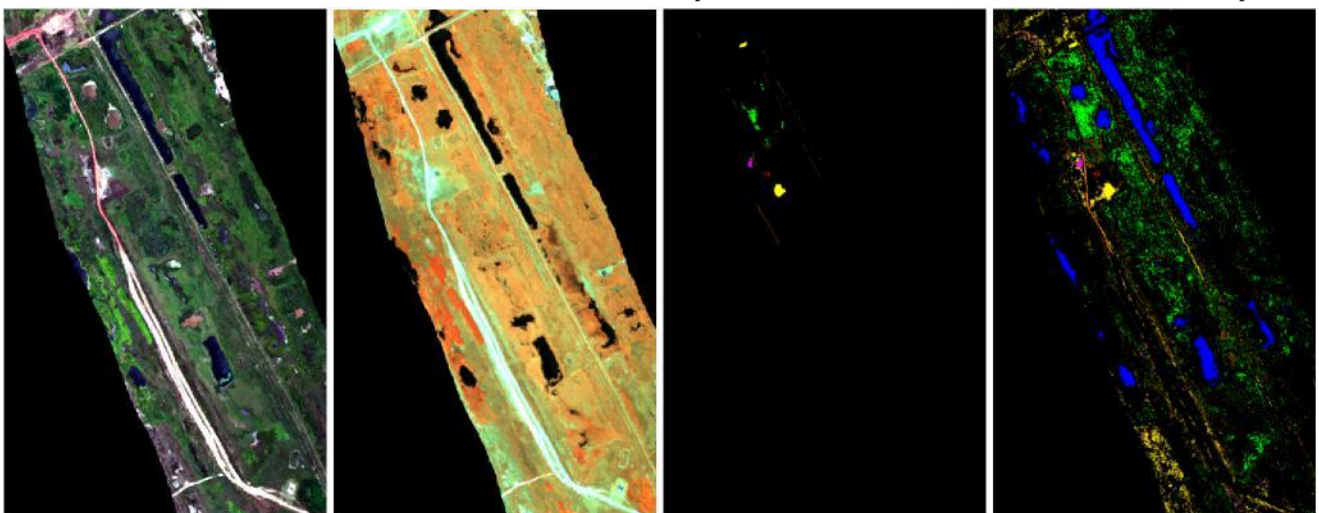

Figure 2 (a) True color image   (b) SWIR image   (c) Ground truth (in situ)   (d) Ground truth (extrapolation)
Legend**:** blue: water, grey: grey sand, orange: laterite, red: concrete, purple: old backwater (north), magenta: old backwater (south), brown: short stressed vegetation and naked soil, yellow: short stressed vegetation, dark green: sparse green vegetation, light green: dense green vegetation.

structural materials (lignin and cellulose) and water content induce a response in the short-wave infrared (SWIR) domain (Sils et al., 2002), (Ustin et al., 2009), (Cheng et al., 2006).
Therefore, a first approach consists in identifying three main classes on the hyperspectral image: "water", "bare soil/man-made structure", "healthy vegetation". The remaining unclassified pixels are assumed to be the stressed vegetation that might indicate a contaminated soil. We rely on specific absorption indexes based on wavelengths situated in optical atmospheric windows. Although this requires expert knowledge, it can be done without any pixel annotation.

Healthy vegetation is characterized using a combination of the biomass predictor NDVI (Normalized Difference Vegetation Index) (Rouse, 1974), (Thenkabail et al., 2000), the carotenoid pigment index PRI (Photochemical Reflectance Index) (Gamon et al., 1992) and SAVI (Soil Adjusted Vegetation Index) (Huete, 1998). PRI is sensitive to changes in carotenoid pigments in live foliage and uses reflectance at 531 and 570 nm wavelengths. The narrow-band version of the NDVI results in the combination of NIR (830 nm) to red (640 nm) reflectances. It is considered as a good predictor of wet and dry green biomass and reduces sensitivity to non-photosynthetic vegetation. The predominant vegetation on the site is herbaceous with moderate leaf area index value and there is no risk of NDVI saturation. SAVI is used to improve canopy structure estimates and minimize the impact of the atmosphere and substrate. It uses the spectral bands at 700, 780 and 900 nm.

The bare soil/man-made structure is classified using the revised OSAVI (WU et al., 2008) to separate bare soil and arid vegetation, and NDBI (Normalized Difference Built-Up Index) (Zha et al., 2003) and UI (Urban Index) (Kawamura et al., 1996).
The revised OSAVI index is used to separate the pixels between the bare soil and the arid vegetation. It exploits the spectral band at 705 nm and 750 nm. NDBI and UI have been employed in various studies for mapping the built-up and bare land in urban areas. NDBI and UI are applied to narrow spectral bands even if there were specified for satellite with broad spectral bands. NDBI is based on the spectral bands at 855 and 1635 nm. UI uses the bands at 830 and 2210 nm.

Finally, we detect water bodies pixels by combining the HDWI (New Hyperspectral Difference Water Index) (Xie et al., 2014) and NDWI (Normalized Difference Water Index) indices. NDWI was initially specified to delineate open water features using the green and NIR bands of Landsat TM. HDWI is constructed to increase the contrast between water and other dark surfaces like shadowed regions by taking advantage of the differences in the spectral amplitudes, particularly in the red and the NIR regions of the spectra.

The classified image using the index combination is shown in Fig. 3. The unclassified pixels (in black) are considered as stressed vegetation. Stressed vegetation areas are located near the factory, along the pipeline and around the pit. These results are consistent with what was observed in the field. Notably, the vegetation that grew over the well mudpit is distinctively detected. This mudpit is indeed surrounded by an earth dam made of sand, which is consolidated by adding heavy hydrocarbons, as a sort of tarmac. In addition, the vegetation on top the dam is regularly mawn.
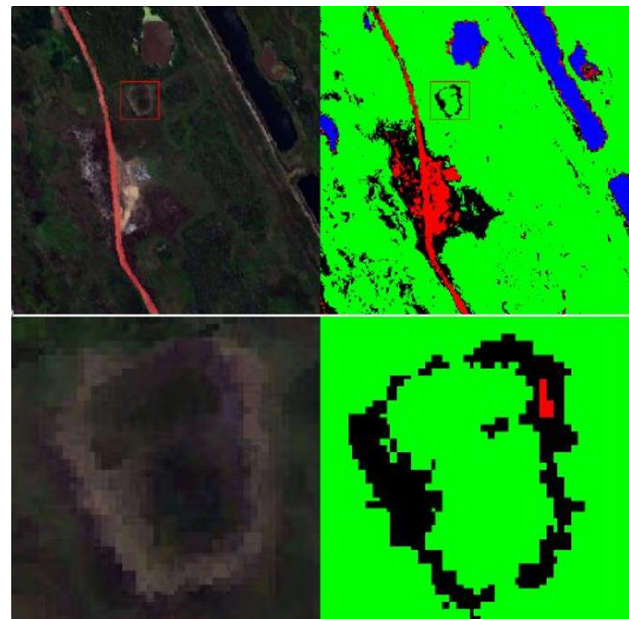

Figure 3 True colour image (left) and classification results (right). Legend: red: bare soil/man-made structure, blue: water, green: healthy vegetation, black: stressed vegetation.

Further to the North-East (Fig.4), the vegetation on top of the ridge covering a pipe also appears as stressed. In the field it appeared as a different vegetation than the one observed on both side of the pipe as seen on Fig. 5.
The ground soil is very sandy and permeable. As a result, the humidity level at the top of the ridge was less than 1%, which is very low. This very low water content probably explains the sparse vegetation and the presence of different plant species than in the more humid flanks of the pipe.
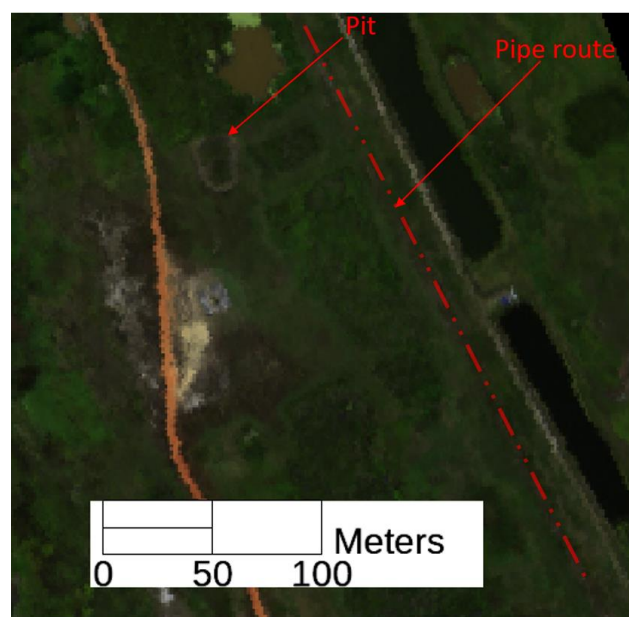

Figure 4: Respective locations of mudpit and pipe

4.2 Unmixing and endmember classification

Based on these promising results, we try to investigate further how much information can be extracted from the hyperspectral image. Especially, we aim to identify different kinds of materials and vegetations and their distribution in the scene.

In order to give an overview of the polluted area in a context wherein no prior information about the scene available, we use an original automatic classification called UM-SVM, combining unmixing and SVM classification (support Vector Machine) (Achard et Al., 2018) .
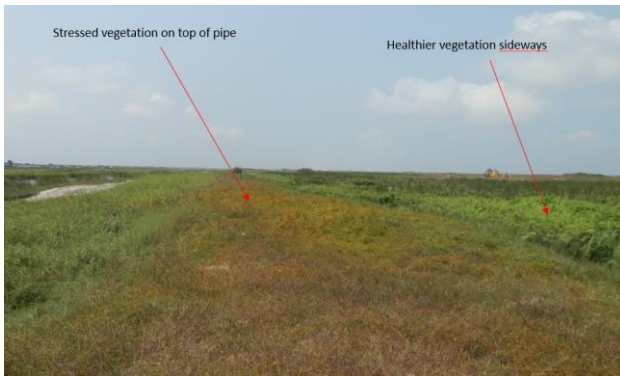

Figure 5: South-East ward looking view showing sparse drier vegetation on top of pipe compared to the vegetation on the sides. Humidity on top on pipe is less than 1%.

First, we apply an automatic endmembers (EM) extraction (AEE) process using a deterministic approach based on OSP (Orthogonal Subspace projection) [20]. Abundances are then computed with fully constrained unmixing method. We extract learning samples from the abundance maps and use them to train a SVM for classification.
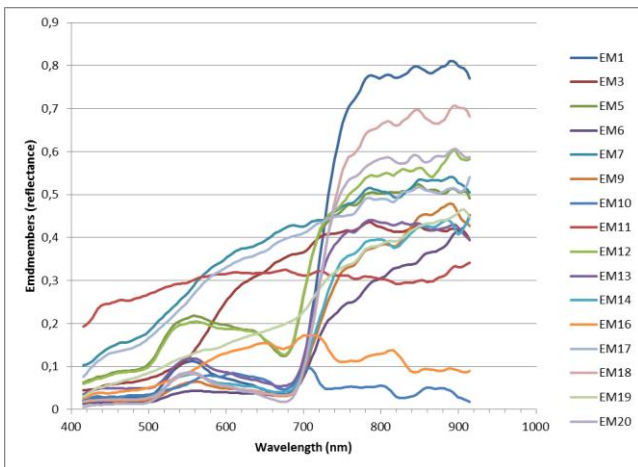

Figure 6: Endmembers obtained with OSP

Determining the number of EM in hyperspectral unmixing is a complex issue that has motivated several works (Boucas-Dias, Nascimento, 2008), (Chang, Du, 2004). Here, we benefit from OSP AEE as increasing the EM numbers do not change the already extracted EM. Thus, we are less sensitive to the initial number of EM. When selecting learning samples from abundance maps, some classes are rejected based on a criterion of too small set of samples with significant abundances. This step, also makes the method less sensitive to the chosen EM number, seeing that if enough EM are chosen, the smallest sets will be automatically filtered out. In this study, we initially

choose 20 endmembers, 16 of which are finally retained. The corresponding endmembers are shown in Fig. 6.

Finally, we apply an SVM classification on the whole image using a polynomial kernel of degree 3. The result is illustrated on Fig. 7. We can then manually map the endmembers with the classes of interest. This approach was already tested by Achard et Al. in a similar environment (Achard, 2018) and giving good results. Here, the classification results are consistent with the extended ground truth although the "stressed vegetation" covers a very wide area. This confirms that the class taxonomy is consistent with the actual pure materials present in the hyperspectral image, but questions where to set the limit of what is defined as stressed vs healthy vegetation.
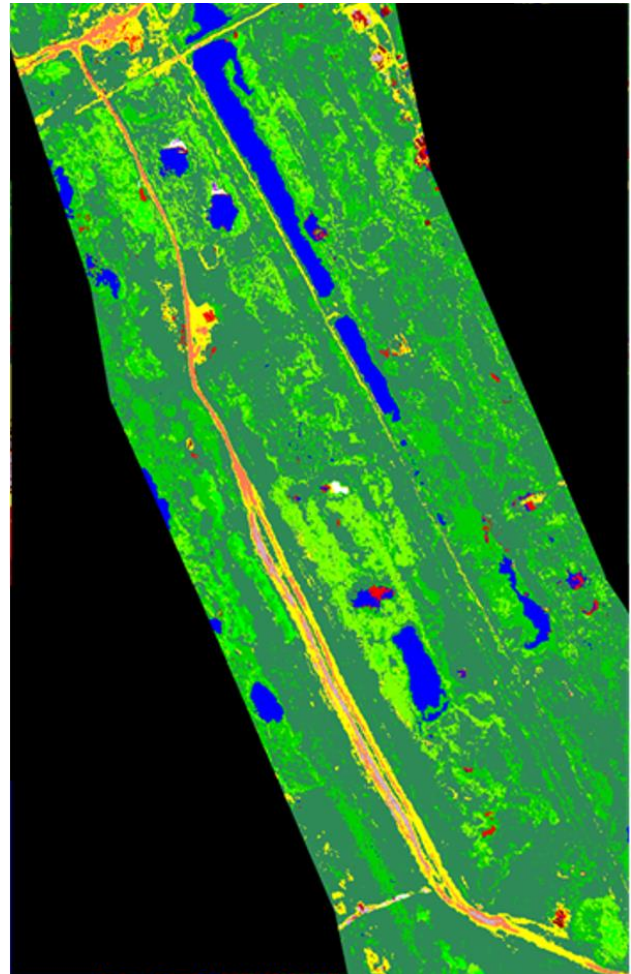

Figure 7: Classification map obtained using UM-SVM. Legend: Blue: water (EM10), Yellow: Bare soil/very sparse vegetation (EM19), Very dark green: Stressed vegetation (EM6), Dark green: Green vegetation (EM13,14), Green: Green vegetation (EM 1,18,20), Light green: Green vegetation (EM3). Other colors: man-made and bare soils (EM 5, 7, 11, 12, 16, 17).

4.3 Deep learning-based classification

Finally, we investigate the use of deep neural networks for automated classification using all the classes identified by the in-situ campaign. The goal is to see whether it is feasible to automatically parse the data using a deep network to automatically identify the materials of the scene, especially when dealing with future acquisitions. Deep learning has established new state-of-the-art for hyperspectral image

| Method | Water | Laterite | Sand | Backwater | | Concrete | stressed vegetation | | Healthy vegetation | | K |
				North	South		short	mixed w/ soil	Sparse	dense	
SVM	100	99.9	84.3	47.3	44.6	100	66.3	52.8	47.4	78.8	65.0
1D NN	100±0.0	99.9±0.1	84.3±0.5	59.1±0.3	54.6±1.4	93.6±2.6	78.5±0.2	67.9±0.3	63.2±0.6	85.9±0.8	76.2±0.5
3D CNN [3]	100±0.0	100±0.0	90.1±1.2	53.2±2.0	51.1±0.7	98.5±0.9	79.3±0.6	65.5±1.1	65.1±0.8	91.5±0.5	78.2±0.3

Table 2: Classification scores for three ML algorithms

classification using 2D and 3D Convolutional Neural Networks. These approaches use trainable convolutional kernels stacked with non-linear activation functions that operates both in the spectral and spatial dimensions. This allows the network to automatically learn a representation of the data that combine spectral and spatial features suitable for classification. Deep networks are known to be one of the must data-hungry statistical learning model, however the hyperspectral dataset is larger than all of the publicly available annotated hyperspectral datasets such as Pavia or Indian Pines.

We choose to compare three different models for this multi-class classification task: a baseline linear SVM, a 1D 3-layers deep fully connected neural network and the 3D CNN proposed in [3]. This covers both spectral and spatial-spectral approaches. We train each network on the original ground truth (Fig. 2c) and we measure the agreement score with the extrapolated annotations from Section 3 (Fig. 2d). Results are detailed in Table 2 and show that neural networks can significantly improve over traditional SVM approaches. Moreover, it stresses out that spectral approaches can be very effective but also improve performance for classes that possess specific spatial properties, such as concrete structures and homogeneous dense vegetation patches. Discriminating between types of soil is challenging (e.g. separating the dried backwater from short vegetation mixed with soil) and could be addressed separately to improve the accuracy.

5. CONCLUSION

This work presents a real-world use case of hyperspectral imagery and its applications in material and vegetation classification. Especially, we acquire a large dataset over a production area with over 50 years of production history in order to assess the feasibility of performing hydrocarbon detection using vegetation health as a proxy.

Based on sparse in situ measurements, we develop a traditional index based classification, an unsupervised UM-SVM classification and two machine learning strategies to extract knowledge from the hyperspectral data. First, we show that hand-crafted reflectance indices can be used for classification of healthy and stressed vegetation. In this case study, this approach is quite efficient. Then, we show how an unsupervised un-mixing can be used to bootstrap an SVM for land cover classification. Land cover classification is god but with probably a too large stressed vegetation class. Finally, we show that deep neural networks can effectively leverage large amount of hyperspectral data for land cover classification and vegetation health monitoring. If we are looking at precision metrics neural networks are performing well.

This case study revealed that they were several types of vegetation stresses; hydrocarbon related stress was observed around the mudpit, hydric stress even in this tropical environment on top of pipes, and stressed vegetation at the bottom of dried near shore lagoons. The classification map is therefore not a map of oil impacted areas, but it will help reduce significantly the areas to check on the ground to identify impacted areas.

REFERENCES

Achard V., Fabre S., Alakian A., Dubucq D., Déliot P., 2018 : " Direct or indirect on-shore hydrocarbon detection methods applied to hyperspectral data in tropical area", SPIE Remote Sensing, Sep 2018, BERLIN, Germany. 10.1117/12.2325097. hal-01969425

Arellano P., Tansey K., Balzter H., *et al.*, 2015: "Detecting the effects of hydrocarbon pollution in the Amazon for-est using hyperspectral satellite images," *Environmental Pollution*, vol. 205, pp. 225–239, Oct. 2015.

Ben Hamida A., Benoit A., Lambert P., *et al.*, 2016: "Deep Learning Approach for Remote Sensing Image Analysis," in *Big Data from Space*, Santa Cruz de Tenerife, Spain, Mar. 2016, p. 133.

Berk A., Anderson G. P., Bernstein L. S., *et al.*, 1999: "MOD-TRAN4 radiative transfer modeling for atmospheric correction," vol. 3756, International Society for Optics and Photonics, Oct. 1999, pp. 348–354.

Bioucas-Dias J. M., Nascimento J. M. P., 2008: "Hyper-spectral subspace identification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 8, pp. 2435–2445, Aug. 2008.

Chang C.-I., Du Q., 2004: "Estimation of number of spectrally distinct signal sources in hyperspectral im-agery," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 3, pp. 608–619, Mar. 2004.

Chang C. I., Xiong W., Chen H. M., *et al.*, 2011: "Maximum orthogonal subspace projection approach to estimating the number of spectral signal sources in hyperspectral imagery," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 504–520, Jun. 2011.

Cheng Y.-B., Zarco-Tejada P. J., Riaño D., *et al.*, 2006: "Estimating vegetation water content with hyperspectral data for different canopy scenarios: Relationships be-tween AVIRIS and MODIS indexes," *Remote Sens. Environ.*, vol. 105, no. 4, pp. 354–366, Dec. 2006.

Gamon J. A., Peñuelas J., Field C. B., 1992: "A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency," *Remote Sens. Environ.*, vol. 41, no. 1, pp. 35–44, Jul. 1992.

Huete A. R.,1988: "A soil-adjusted vegetation index (SAVI)," *Remote Sens. Environ.*, vol. 25, no. 3, pp. 295–309, Aug. 1988.

Kawamura M., Jayamana S., Tsujiko Y.,1996: "Relation between social and environmental conditions in Colombo, Sri Lanka and the Urban Index estimated by satellite remote sensing data.", *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci*, vol. 31, no. B7, pp. 321–326, 1996.

Lassalle G., Credoz A., Hédacq R., *et al.*,2018: "Assessing Soil Contamination Due to Oil and Gas Production Using Vegetation Hyperspectral Reflectance,", *Environmental Science & Technology*, 2018 52 (4), 1756-1764. DOI: 10.1021/acs.est.7b04618

Miesch C., Poutier L., Achard V., *et al.*,2005: "Direct and inverse radiative transfer solutions for visible and near-infrared hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 7, pp. 1552–1562, Jul. 2005.

Noomen M. F., van der Werff H. M. A., van der Meer F. D., 2012: "Spectral and spatial indicators of botanical changes caused by long-term hydrocarbon seepage," *Ecological Informatics*, vol. 8, pp. 55–64, Mar. 2012.

Plyer A., Colin-Koeniguer E., Weissgerber F.,2015: "A new coregistration algorithm for recent applications on urban sar images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2198–2202, Nov. 2015.

Rouse J. W.,1974: "Monitoring vegetation systems in the Great Plains with ERTS," in *3rd Earth Resource Technology Satellite (ERTS) Symposium*, vol. 1, Jan. 1974, pp. 48–62.

Sims D. A., Gamon J. A., 2002: "Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages," *Remote Sens. Environ.*, vol. 81, no. 2, pp. 337–354, Aug. 2002.

Thenkabail P., Smith R., De Pauw E.,2000: "Hyper-spectral vegetation indices for determining agricultural crop characteristics," *Remote Sens. Environ.*, vol. 71, pp. 158–182, 2000.

Ustin S. L., Gitelson A. A., Jacquemoud S., *et al.*,2009: "Retrieval of foliar information about plant pigment sys-tems from high resolution spectroscopy," *Remote Sens. Environ.*, vol. 113, S67–S77, Sep. 2009.

Xie H., Luo X., Xu X., *et al.*, 2014: "New hyperspectral difference water index for the extraction of urban water bodies by the use of airborne hyperspectral images," *J. Appl. Remote Sens*, vol. 8, no. 1, pp. 85–98, Jul. 2014.

Wu C., Niu Z., Tang Q., *et al.*,2008: "Estimating chlorophyll content from hyperspectral vegetation indices: Model-ing and validation," *Agricultural and Forest Meteorol-ogy*, vol. 148, no. 8, pp. 1230–1241, Jul. 2008.

Zha Y., Gao J., Ni S.,2003: "Use of normalized differ-ence built-up index in automatically mapping urban ar-eas from TM imagery," *International Journal of Remote Sensing*, vol. 24, pp. 583–594, Feb. 2003.