MAPPING OIL SPILLS ON SEA SURFACE FROM SENTINEL 2 IMAGES USING PRINCIPAL COMPONENTS AND CATEGORICAL BOOSTING

R.J.L. Argamosa ^{a*}, A.C. Blanco ^{a,b}, R. B. Reyes ^{a,b}

^a Training Center for Applied Geodesy and Photogrammetry, University of the Philippines, Diliman, 1001, Philippines – rlargamosa@up.edu.ph, acblanco@upd.edu.ph, rbreyes3@up.edu.ph

^b Department of Geodetic Engineering, University of the Philippines, Diliman, 1001, Philippines

Commission 4, WG 7

KEY WORDS: Principal component analysis, Water remote sensing, Homogeneity, Image texture, Gradient Boosting

ABSTRACT

A large oil spill in Iloilo Straight that occurred on July 3, 2020, as well as a possible deliberate, small but frequent oil spill and surfactant contamination in Manila Bay, were mapped. The method employs the Sentinel 2-1C image, which is transformed into principal components to reveal the presence of oil spills and possibly surfactants. Additionally, a gradient boosting algorithm was trained to discriminate between pixels that were contaminated with oil and those that were not. The multi-band image with three principal components with a 99% cumulative explained variance ratio highlights the occurrence of an oil spill in Iloilo Straight. Further, the classified image produced by pixel-based classification clearly distinguishes between water and oil pixels in the said area. The methodology was applied to a Sentinel 2-1C image of Manila Bay, with pixels observed/identified as oil and classified as well. The highest density of supposedly oil-contaminated pixels (large or small but frequent) was observed on the eastern side of Manila Bay (Bataan). While there were no documented oil spills concurrent to the satellite image used, historical reports on the area indicate that the likelihood of an oil spill is extremely high due to the massive amount of shipping activity. Pixels supposedly contaminated by oil spills also occur in areas near ports where oil spills could occur as a result of ship operations. Pixels with the same properties as oil contamination are also visible in areas adjacent to fishponds and aquaculture, where phytoplankton and fish contribute to surfactant contamination.

1. INTRODUCTION

An oil spill accident results in the rapid leakage of a large amount of oil. Oil slicks caused by accidents float to the surface of the sea, wreaking havoc on the marine environment (Zhao et al., 2018). Some oil pollution is caused not by ship collisions, but by routine ship operations such as tank cleaning and engine effluent discharges. In 2010, the Deepwater Horizon platform oil spill in the Gulf of Mexico raised major environmental concerns (Leifer et al., 2012; Garcia-Pineda et al., 2013 as cited by (Alpers et al., 2017). Contrary to the 1983 MARSPOOL 73/78 International Convention for the Prevention of Pollution from Ships, large amounts of mineral oil are still illegally discharged into the sea. Tank washing and engine effluent discharges cause the majority of anthropogenic oil pollution at sea (sludge) (Alpers et al., 2017). Offshore oil platforms and refineries are anthropogenic sources of pollution (Alpers et al., 2017). On the other hand, biosurfactants are amphiphilic compounds produced by bacteria that can help break down oil. These surfactants form slicks on the sea surface, altering the physical properties of the near-surface layer of the ocean by damping short gravity-capillary waves and suppressing turbulence structures (Parks et al., 2020). Biota in the water column, including phytoplankton and fish, secrete surfactants below the wave-stirred water layer. Only in the form of biogenic surface films can biogenic surface films exist on the sea surface. They are monolayers, with a thickness of only one molecular layer (typically 2.4-2.7 nm). They are made up of surface-active material (or surfactants) secreted by biota in the water column, such as phytoplankton and fish, and are typically located below the wave-stirred water layer (Wurl et al., 2016; Kurata et al., 2016 as cited by Alpers et al., 2017)

When it comes to spectral properties of oil in water surfaces, Li et al., (2012 studied the spectral range of 550nm to 750nm between

crude oil and other objects to detect crude oil floating on the sea surface. Using Sentinel 2 image, the best results for detecting pixels contaminated with oil spills were using the bands difference between 660 and 560 nm, division of 660 and 560 nm, and division of 825 and 560 nm, normalized by 480 (Taravat and Del Frate, 2012 as cited by Kolokoussis and Karathanassi, 2018). In near real-time data, the wavelength region around 344.51 nm is most suitable for oil type discrimination. This wavelength region has constant signature ratios between oil reflectance and pure water reflectance for each oil type. Oil spill reflectance values and thickness are highly correlated in a specific wavelength region. These regions are found between 474 and 917 nm. For each oil type, there is a high correlation between reflectance values and age. These regions are found between 576 and 919 nm (Andreou et al., 2011). Several literatures indicate that oil film has a reflectance in the 400-700 nm wavelength range, in contrast to background seawater (Rajendran et al., 2021).

Oil pollution of the sea surface is a major environmental concern. Monitoring accidental or illegal oil discharges is critical to reducing marine pollution. However, most research on oil detection algorithms focused on large oil spills, while deliberate and frequent small oil spills were rarely studied.

The goal of the study is to detect pixels contaminated by oil spills or by possible surfactants. Because both oil and surfactants dampen the gravity-capillary waves, the presence of the latter will be detected in effect. The study has the following objectives: (1) transform Sentinel 2 bands and other derived layers into principal components, and (2) create an RGB multiband composite to highlight oil contaminated pixels (3) using a gradient boosting algorithm, classify the uncontaminated and oil contaminated water pixels, and (4) apply the developed methodology to the Sentinel 2 image of Manila Bay.

2. STUDY AREA AND DATA

2.1. Study area

This study considered two areas. The first is in Iloilo Straight, where a large accidental oil spill has been reported. This was where oil-contaminated pixels were selected as regions of interest and then trained using a gradient boosting classifier with principal components as variables. The second is in Manila Bay, where the first area's data transformer and trained classifier will be implemented. While there was no known large oil spill in Manila Bay at the time of Sentinel 2 acquisition that were used in this study, pixels detected as oil contaminations were considered to be surfactants because surfactants, like oil, dampen capillary waves. Detected oil-like pixels could also be caused by operational and unintentional oil spills by small boats and ships.

The Province of Iloilo is located on Panay Island's southern and northeastern coasts. It is bounded on the north by the Province of Capiz and the Jintotolo Channel; on the south by the Panay Gulf and the Iloilo Strait; on the east by the Visayan Sea and the Guimaras Strait; and on the west by the Province of Antique (Province of Iloilo, 2014). The Iloilo Strait connects the Panay and Guimaras Islands and connects the Panay Gulf to the Guimaras Strait. On July 3, 2020, around 48,000 liters of oil spilled into waters off Iloilo City on Friday following an explosion at a power barge owned by AC Energy corporation and the estimated spill affected an area of 1,200 square meters (CNN Philippines, 2020).

Manila Bay is an estuary that is partially enclosed. The bay is located on the southwest coast of Luzon Island, one of the Philippines' major islands. It is located between the longitudes of 120°28 and 121°15 East and the latitudes of 14°16 and 15° North. It has a roughly 190-kilometer-long coastline and a surface area of approximately 1,800 square kilometers. It is bounded by the National Capital Region's coastal cities and municipalities, as well as the coastal provinces of Bataan, Pampanga, and Bulacan in Region 3, and Cavite in Region 4. The Pasig and Pampanga River basins are the two primary contributory areas. Discharges vary significantly by season and year, with the highest input occurring in August and the lowest in April (PEMSEA, 2001). Agriculture, forestry, and fishing are the main economic activities in the bay's catchments. Industrial activities include manufacturing, mining, and quarrying. Food and beverage, chemical, pharmaceutical, petrochemical, and electronic industries are major manufacturers. The fishing trade is heavily reliant on both local and distant fishing grounds. The shipping industry transports passengers, oil, and various containers. Activities like reclamation and construction can impact habitats and contribute to suspended materials in the bay. Agriculture and forestry, especially in river catchment areas, can contribute to agrochemical, agricultural waste, and soil erosion pollution. Manila Bay is drained by a 17,000 km² watershed with 26 catchment areas. The Pasig and Pampanga River basins contribute significantly. Most rivers in Pampanga, Bulacan, and Nueva Ecija drain into the Pampanga (PEMSEA, 2004). Freshwater inflow is estimated at 25 km3/year, but this is likely an overestimate. Seasonal and annual variations in discharges are notable, with August being the highest and April the lowest. Freshwater retention time in the bay varies from two weeks to one month depending on the season (PEMSEA, 2004). The tidal range is 1.2 m during spring tide and 0.4 m during neap tide. Weather patterns, especially in shallow water, are influenced by seasonal and diurnal winds. During the wet season, the water

column's salinity increases from surface to bottom. Median salinity is between 30 and 35%, slightly less than the open ocean, with levels dropping during the rainy season. A slight seasonal and temporal variation of around 30° C (PEMSEA, 2004). Mangroves and fisheries are among Manila Bay's natural resources. Below is a map of the two study areas.



Figure 1. Location map of Manila Bay (left) and Iloilo Straight (top right). Source: OpenStreetMap

2.2. Data

Sentinel 2's surface reflectance product was not used because the atmospheric correction algorithm removed water reflectance values; instead, both sites used Sentinel 2-1C or top of atmosphere reflectance. Both datasets were processed prior to being downloaded from Google Earth Engine (Gorelick et al., 2017). Cloud masking was applied to both images using the bitmask layer and land masked using Normalized Difference Water Index (NDWI). For Iloilo Straight, a single image was chosen which was July 8, 2020, or five days after the oil spill occurred, while for January 2020 data, a monthly composite with a cloudy pixel percentage of less than 5% was used. A monthly composite image of Manila Bay was created to create a seamless and nearly gapfree image. Figure 2 shows the true color and false color composites of Iloilo Straight.



Figure 2. Sentinel-2 true color (left) and false color composite (right) of Iloilo straight (top) and Manila Bay (bottom). The arrows indicate oil-contaminated water because of the large oil spill that occurred accidentally. Sentinel 2 sensing date: July 8, 2020 for Iloilo Straight and January 2020 composite for Manila Bay

3. METHODOLOGY

3.1. Data reduction

Principal component analysis (PCA) is a multivariate statistical technique for identifying uncorrelated linear combinations of variables with smaller variances (Loughlin, 1991). PCA is an unsupervised linear transformation technique used to reduce dimensionality. PCA finds patterns in data by correlating features and transforming high-dimensional data into a subspace of equal or fewer dimensions. Given that the new feature axes are orthogonal to each other, the orthogonal axes (principal components) of the new subspace are the directions of maximum variance. A d x k- dimensional transformation matrix W will be constructed to map a sample vector x into a k-dimensional feature subspace that has fewer dimensions that the original d-dimensional features space:

$$\boldsymbol{x} = [x_1, x_2, \dots \, x_d], \, \boldsymbol{x} \in R^d, \tag{1}$$

$$xW, W \in \mathbb{R}^{dxk}$$

$$\mathbf{z} = [z, x_2, \dots x_k], \mathbf{z} \in \mathbb{R}^k \tag{3}$$

Transforming the original d- dimensional into k-dimensional subspace will result in the first principal component that has the highest variance, and all subsequent principal components have the highest variance because they are uncorrelated (Raschka, 2016). When variables are multispectral image channels, both the spatial abundance of various surface materials and image statistics influence the ordering of principal components(Loughlin, 1991). In this case, principal components were used to compress the layers and then create a composite. The number of principal components to use was based on explained variance percentage.

Three components were chosen to be included in the composite. The multiband composite was used to highlight water pixels contaminated with oil when visualized in an RGB composite image. PCA was implemented in Scikit-learn library (Varoquaux et al., 2015) in Python using the *fit* function of the PCA module after scaling the values to 0-1. The PCA transformer was then saved that was used to images Layers that were transformed into principal components were the visible range, near infrared, and short-wave infrared. Moreover, the ratio of blue and short-wave was also added since it was found out through visual inspection and the study conducted by Kolokoussis and Karathanassi (2018) shows that oil contaminated pixels can be distinguishable from the ratio of the said bands. This is because longer wavelengths, like shortwave, have less reflectance, resulting in darker pixels.

The spatial relationship between gray levels/DNs in an image contains textural information. Texture reflects image properties like smoothness, coarseness, and regularity. The texture context is described by statistical, structural, and spectral principles. Statistical techniques characterize textures as smooth, coarse, grainy, and other qualitative measures. The spatial relation is the covariance of pixel values with distance and direction. These matrices can be used to extract information from images (C). Homogeneity is a measure of C uniformity, and it is high if most elements are on the main diagonal (Navulur, 2007). Thus, pixels with capillary waves dampened by oil/surfactants will have high homogeneity due to less reflectance and therefore less variation in values. Table 1 and Figure 3 lists and shows, respectively the layers that were used for PCA.

Layers	Wavelength/Formula
Blue	496.6nm (S2A) / 492.1nm (S2B)
Green	560nm (S2A) / 559nm (S2B)
Red	664.5nm (S2A) / 665nm (S2B)
Near-infrared	835.1nm (S2A) / 833nm (S2B)
Shortwave infrared	1613.7nm (S2A) / 1610.4nm (S2B)
B/SWIR	Blue/SWIR
Homogeneity	Homogeneity = $\sum_i \sum_j C_{ij}^2$ where $C = \frac{B}{aver}$
	SWIR SWIR
Table 1 I avers that were compressed into principal	



3.2. Training CatBoost

Boosting adds new models to the ensemble sequentially. Each iteration trains a new weak base-learner model based on the ensemble's error. The learning procedure continuously fits new models to improve the response variable estimate. This algorithm's main idea is to build new base-learners that are maximally correlated with the ensemble's negative gradient loss function (Natekin and Knoll, 2013). A powerful gradient boosting algorithm is CatBoost (Categorical Boosting). It is a decision tree gradient boosting algorithm developed by company Yandex. It achieves cutting-edge results without the need for extensive data training. It reduces the need for hyper-parameter tuning and overfitting, leading to more generalized models (Prokhorenkova et al., 2018). To optimize models, the Optuna module was used where 100 trials were tested to find the optimal set of hyper parameters. Optuna is a software framework for automatic hyperparameter optimization (Akiba et al., 2019). Regions of interest were selected for oil and non-oil pixels as shown in Figure 4. Selection for each class were based on visual interpretation where dark pixels correspond to oil slicks. Binary classification was done using the CatBoost classifier. The overall workflow used for mapping oil spill is shown in Figure 5.



Figure 3. Layers that were transformed to principal components. The low reflectance of oil-contaminated pixels in the Sentinel 2

bands can be seen in the visible, near-infrared, and shortwaveinfrared. High values can be observed for the blue-to-shortwaveinfrared ratio and its Homogeneity texture.



Figure 4. Location of pixels selected as oil and non-oil.



Figure 5. The overall workflow for mapping oil spill

4. **RESULTS**

4.1. PCA transformation and classification result

The variance ratio explained below shows that the first component alone accounts for 90% of the variance and with the second and third components combined account for 99%. Thus, the seven layers (visible, near-infrared, shortwave infrared, B/SWIR, and homogeneity) can be reduced to three. Using the tuned CatBoost classifier model, the classification result achieved a 99 percent accuracy on test data (accuracy is computed as the total number of pixels correctly classified divided by total number of pixels). This demonstrates the effectiveness of classifying oil spills in Iloilo Straight using the principal components of the selected layers. The raster representation of principal components is shown in Figure 6 and Figure 7 is the RGB composite of the 3 PCAs.



Figure 6. Principal components represented as a raster in Iloilo Straight. The explained variance ratios for PC1, PC2, and PC3 were 90%, 10%, and 1%, respectively.



Figure 7. Multi-band composite consisting of PC1, PC2, and PC3 arranged in the following: Red: PC2, Green: PC1, and Blue: PC3. On the right is the result of pixel-based classification. The PCA multiband composite contains bright red pixels that correspond to the oil classified pixels.

The blue-to-shortwave-infrared ratio and homogeneity of the image of Manila Bay were calculated. Following that, principal components were derived using the PCA transformer generated from Iloilo Straight. Figure 8 illustrates the multi-band composite and the result of the pixel-based classification.



Figure 8. A multi-band composite for Manila Bay is created (left) after application of the PCA transformer developed for mapping oil spill in Iloilo Straight. The red coloration corresponds to oil-like features. A binary image showing the oil contaminated and uncontaminated areas is shown in the right.



Figure 9. Zoomed in images of the areas in Manila Bay with possible oil slicks. Left is Metro Manila and Cavite; Middle image is Bataan and right image is Pampanga where aquaculture where widespread aquaculture practices exist

The presence of oil-like features in the westside portion of Manila Bay (Bataan) may indicate small but frequent oil spills. While there have been no reports of large and accidental oil spills in the area during the acquisition of Sentinel 2 images, it has been reported historically that the likelihood of an oil spill in the Bataan (middle image in Figure 9) coast is very high due to the massive amount of shipping activity. Additionally, numerous fishing boats operating in the area can cause operational and unintentional oil spills (PEMSEA, 2017). In 1992 and 1993, a sample from Amo, Mariveles, Bataan, had the highest level of oil and grease. Oil refineries in Mariveles and Limay, Bataan, may explain the observations. Globally, oil is thought to enter the marine environment via land-based sources such as refineries, municipal and institutional wastes, and urban runoff (GESAMP, 1993 cited in MPP-EAS, 1999b). The relative contribution of land-based and sea-based sources varies depending on the site's circumstances. Oil spills from land and sea sources contribute to the oil in Manila Bay (PEMSEA, 2004).

Pixels with the same properties as oil contamination are also visible in areas adjacent to fishponds and aquaculture (northeast side, rightmost image in Figure 9), where phytoplankton and fish contribute to surfactant contamination (Alpers et al., 2017). Pixels supposedly contaminated by oil spills also occur in areas near ports (Metro Manila and Cavite area, leftmost image in Figure 9) where oil spills could occur as a result of ship operations.

Since ships contribute to possible oil spills, a preprocessed Sentinel 1 GRD composite (reduced using max value of backscatter) of the same month was generated and correlated to the presence of detected ships by the radar satellite. Only images from January to March were generated due to cloud contamination in the following months. The following images are shown below:





Figure 10. Top: January 2020 - Oil/surfactant pixels were found in the same area as increased ship occurrence in Bataan. Middle: February 2020 - Like the previous image, oil/surfactant pixels were found in Metro Manila near increased ship occurrence. Bottom: March 2020 – Unlike the previous two images, the radar image did not contain any formations resembling ships. While no evidence of strong backscatter indicative of ships was found, a

strong backscatter indicates strong winds (Gao et al., 2021), which resulted in the wide distribution of possibly surfactants originating in aquaculture areas.

5. CONCLUSION

This study was able to map events that could dampen capillary waves by transforming the visible, near-infrared, shortwave infrared, the ratio of blue and shortwave infrared (BSWIR), and the homogeneity texture of BSWIR into principal components. Ninety-nine percent of the variance in the multiband distribution was explained by the first three principal components. Through the use of red coloration, the aforementioned multiband demonstrates the presence of oil-like pixels and possibly surfactants in the RGB image, indicating the presence of oil spills in the Iloilo Strait. Additionally, the aforementioned PCA multiband was classified using the CatBoost algorithm (which achieved a 99 percent accuracy in the test data) to distinguish between pixels contaminated with oil/surfactants and uncontaminated pixels. The multiband and classified images both demonstrated the ability to detect oil spills. Additionally, surfactants are mapped in effect because they dampen the capillary waves of water in a manner similar to oil. The resulting PCA data transformer and trained model were then applied to Manila Bay, and the resulting PCA multiband image of the area, as well as the binary image result of the CatBoost model, both contained the same red coloration. While there were no reports of a large oil spill at the time of the image acquisition used in this study, oil may have been present due to ship and small fishing boat operations. Surfactants were also considered, given Manila Bay's abundance of aquaculture areas. Due to COVID-19 restrictions, no ground validation of oil presence was conducted.

The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVI-4/W3-2021 Joint International Conference Geospatial Asia-Europe 2021 and GeoAdvances 2021, 5–6 October 2021, online

ACKNOWLEDGEMENTS

This research was done as part of the Development of Integrated Mapping, Monitoring, and Analytical Network System for Manila Bay and Linked Environments (MapABLE) Project under the Integrated Mapping, Monitoring, Modelling and Management System for Manila Bay and Linked Systems (IM4ManilaBay) (IM4ManilaBay) Program. The Project was implemented by the University of the Philippines Training Center for Applied Geodesy and Photogrammetry (TCAGP), through the support of the Department of Science and Technology (DOST) of the Republic of the Philippines and the Philippine Council for Industry, Energy, and Emerging Technology Research and Development (PCIEERD).

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna : A Next-generation Hyperparameter Optimization Framework 2623–2631.

Alpers, W., Holt, B., Zeng, K., 2017. Remote Sensing of Environment Oil spill detection by imaging radars : Challenges and pitfalls. Remote Sens. Environ. 201, 133–147. https://doi.org/10.1016/j.rse.2017.09.002

Andreou, C., Karathanassi, V., Kolokoussis, P., 2011. Spectral library for oil types. 34th Int. Symp. Remote Sens. Environ. - GEOSS Era Towar. Oper. Environ. Monit.

Gao, Y., Sun, J., Zhang, J., Guan, C., 2021. Extreme wind speeds retrieval using sentinel-1 IW mode SAR data. Remote Sens. 13, 1–14. https://doi.org/10.3390/rs13101867

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

ILOILO, P.O., 2014. Annual Provincial Profile 2014.

Kolokoussis, P., Karathanassi, V., 2018. Oil spill detection and mapping using sentinel 2 imagery. J. Mar. Sci. Eng. 6. https://doi.org/10.3390/jmse6010004

Li, Q., Lu, L., Zhang, B., Tong, Q., 2012. Oil Slope Index: An algorithm for crude oil spill detection with imaging spectroscopy. Proc. 2nd Int. Work. Earth Obs. Remote Sens. Appl. EORSA 2012 46–49. https://doi.org/10.1109/EORSA.2012.6261132

Loughlin, W.P., 1991. Principal component analysis for alteration mapping. Photogramm. Eng. Remote Sens. 57, 1163–1169.

Natekin, A., Knoll, A., 2013. Gradient boosting machines, atutorial.Front.Neurorobot.7.https://doi.org/10.3389/fnbot.2013.00021

Navulur, K., 2007. Multispectral Image Analysis Using the Object-Oriented Paradigm.

Parks, G., Dean, C.W., Kluge, J.A., Soloviev, A. V., Shivji, M., Tartar, A., Howe, K.L., Lehner, S., Schwarz, E., Shen, H., Perrie, W., Schuler, P., 2020. Analysis of surfactant-associated bacteria in the sea surface microlayer using deoxyribonucleic acid sequencing and synthetic aperture radar. Int. J. Remote Sens. 41, 3886–3901. https://doi.org/10.1080/01431161.2019.1708508

PEMSEA, 2017. State of the Coasts Report Province of Bataan.

PEMSEA, 2004. Manila Bay : Refined Risk Assessment.

PEMSEA, 2001. Manila Bay Coastal Strategy 121.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. Catboost: Unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst. 2018-Decem, 6638–6648.

Rajendran, S., Vethamony, P., Sadooni, F.N., Al-Kuwari, H.A.S., Al-Khayat, J.A., Seegobin, V.O., Govil, H., Nasir, S., 2021. Detection of Wakashio oil spill off Mauritius using Sentinel-1 and 2 data: Capability of sensors, image transformation methods and mapping. Environ. Pollut. 274, 116618. https://doi.org/10.1016/j.envpol.2021.116618

Raschka, S., 2016. Python Machine Learning.

Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., Mueller, A., 2015. Scikit-learn. GetMobile Mob. Comput. Commun. 19, 29–33. https://doi.org/10.1145/2786984.2786995

Zhao, D., Cheng, X., Zhang, Hongping, Niu, Y., Qi, Y., Zhang, Haitao, 2018. Evaluation of the ability of spectral indices of hydrocarbons and seawater for identifying oil slicks utilizing hyperspectral images. Remote Sens. 10. https://doi.org/10.3390/rs10030421