DATA MINING AND MACHINE LEARNING IN EARTH OBSERVATION – AN APPLICATION FOR TRACKING HISTORICAL ALGAL BLOOMS

Alexandria Dominique Farias and Gongling Sun

International Space University, Strasbourg, France

ABSTRACT

The data produced from Earth Observation (EO) satellites has recently become so abundant that manual processing is sometimes no longer an option for analysis. The main challenges for studying this data are its size, its complex nature, a high barrier to entry, and the availability of datasets used for training data. Because of this, there has been a prominent trend in techniques used to automate this process and host the processing in massive online cloud servers. These processes include data mining (DM) and machine learning (ML). The techniques that will be discussed include: clustering, regression, neural networks, and convolutional neural networks (CNN).

This paper will show how some of these techniques are currently being used in the field of earth observation as well as discuss some of the challenges that are currently being faced. Google Earth Engine (GEE) has been chosen as the tool for this study. GEE is currently able to display 40 years of historical satellite imagery, including publicly available datasets such as Landsat, and Sentinel data from Copernicus.

Using EO data from Landsat and GEE as a processing tool, it is possible to classify and discover historical algal blooms over the period of ten years in the Baltic Sea surrounding the Swedish island of Gotland. This paper will show how these technical advancements including the use of a cloud platform enable the processing and analysis of this data in minutes.

KEYWORDS

Earth Observation, Remote Sensing, Satellite Data, Data Mining, Machine Learning, Google Earth Engine, Algal Blooms, Phytoplankton Bloom, Cyanobacteria

1. INTRODUCTION

Earth observation (EO) has become more prominent in the last decade with more satellites in orbit that are capable of observing the Earth every year. The miniaturization of component parts has also enabled a new generation of CubeSats that are also adding to the data gained from remote sensing (RS). As RS and EO are often used interchangeably, it is worth defining RS as the act of viewing, observing and analysing an object from a given distance. This paper will only be addressing RS data that is observing the Earth, specifically satellite imagery.

The technology providing satellite imagery has improved significantly, with output types ranging from simple traditional photographic images to complex spectral graphs. These developments increase the amount of data that is collected on a daily basis. The data in most cases is so abundant that manual processing is not an option for analysis of all results. As such,

Natarajan Meghanathan et al. (Eds): SIPP, BIGML, DaKM, SOEN, AISC - 2020
pp. 39-58, 2020. © CS & IT-CSCP 2020 DOI: 10.5121/csit.2020.100204

there has been a prominent trend in techniques used to automate this process and host the processing in massive online cloud servers.

These processes include data mining (DM) and machine learning (ML) which will be discussed in this individual report. ML has been emphasized in this study as most methods currently being used in earth observation fall under the heading of 'machine learning'. The types techniques that will be discussed include clustering, regression, neural networks, and convolutional neural networks (CNNs).

This paper will show how some of these techniques are currently being used in the field of earth observation. Some of the challenges of the tools and environments that are currently used will also be discussed. As a practical exploration of these techniques using historical earth observation data, Google Earth Engine (GEE) has been chosen to process and run our scripts on publicly available Landsat RS data catalogues. GEE currently is able to display 40 years of historical satellite imagery and has a built in JavaScript application programming interface (API) that makes geospatial analysis across petabytes of data possible in the Google Cloud [1]. Using this RS data, it is possible to use various DM and ML techniques to classify and discover historical algal blooms in the Baltic Sea surrounding the Swedish island of Gotland.

The organisation of this paper is as follows: Section 2 will discuss data mining patterns and techniques used in EO. Section 3 will look at some of the challenges that are specific to analysing data for EO including discussing training data. In section 4, a 'real-world' example is introduced by looking at historical data in the Baltic Sea and using GEE to identify algal blooms over a period of ten years. Discussions and recommendations are found in section 5 followed by conclusion of the study in section 6.

2. Literature Review of Data Mining Patterns and Techniques in Earth Observation

There are a number of methods that can be used in creating data mining patterns specific to Earth Observation. Amongst these are clustering, regression, neural networks and CNNs. These will be briefly summarized in the following section.

2.1. Clustering

Clustering, according to Berkhin [2], is "a division of data into groups of similar objects." The objects within the cluster groups are similar to each other and not similar to objects from other groups. There are a number of clustering algorithms, amongst these are hierarchical, partitioning and grid-based methods, constraint-based clustering, scalable clustering, and high dimensional data algorithms. Figure 1 below, is an example of clustering.

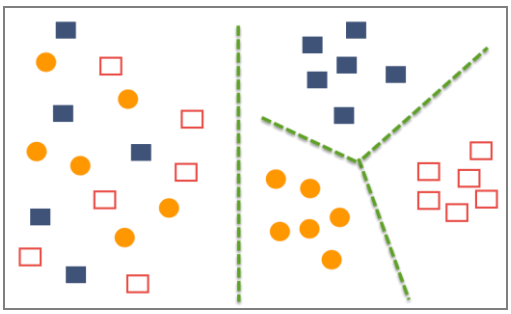

Figure 1. An example of clustering

2.2. Regression

In data mining, as defined by Oracle [3], "Regression is a data mining function that predicts a number." In a regression model, the data set that is used to predict the outcome, has values that are known. The attributes in the data set are called predictors and the outcome is a value which is known as the target. One example of this can be housing cost estimation. The value of the house is the target and the predictors could be attributes as number of rooms, age, location, previous sale costs, etc [3].

Regression can be linear or non-linear. A linear regression is based on the ability to approximate the relationship between the target and the predictors with a straight line. In non-linear regression a relationship is unable to be approximated by a straight line, so a more complex equation has to be defined. In figure 2, a graph with a single predictor is shown for linear regression. The y axis is the target and x is the predictor. The error, also known as the residual, is a measure of the difference between the predicted and the expected value [3].

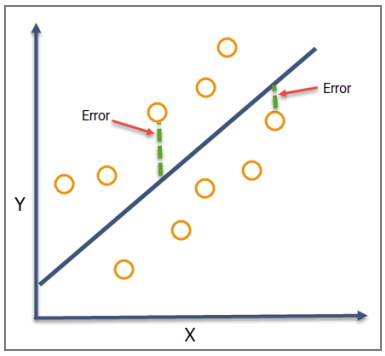

Figure 2. An example of linear regression

2.3. Neural Networks

Neural networks were originally designed to be an analogue of the computation of biological neurons. It has been described by Han, Kamber, and Pei [4], as, "…a set of connected input/output units in which each connection has a weight associated with it." As the network learns, it adjusts the weights of inputs and adjusts accordingly in order to predict classes. This generally involves a long training time and there are criticisms that it is very hard to interpret where the weights come from and what the hidden units are in the network. Neural networks can be described as a 'black box' in that inputs go in and an output is given, but it is unknown what exactly happens inside the box to get to the output. The advantage of neural networks is that they are very tolerant of data that is 'noisy'. They are also very good for classifying patterns that have not been trained and where there is little known about the attributes and classes and the relationships between them [4].

Below in figure 3 can be seen a basic node of a neuron for a neural network. The inputs are given and weights are attached to each. The inputs are then summed and then go through an activation function to determine if the signal should progress further. If the neuron is activated, it is able to contribute to the overall outcome. The outcome can be the function of adding the object to class or participating in a classification.

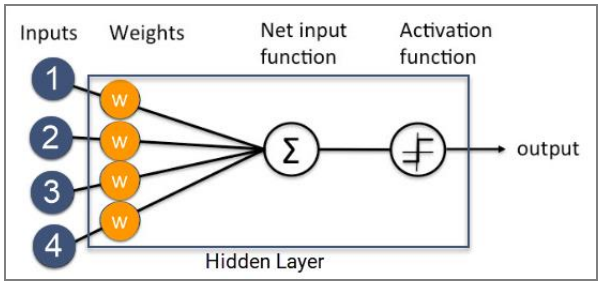

Figure 3: Diagram of a neural network node [5]

Multiple nodes together make a layer and the layer subsequently contributes to the next layer. In this model, there are at least three layers, an input layer, a hidden layer, and an output layer [6]. Neural networks can be classified into three categories, a supervised neural network, an unsupervised neural network and a reinforcement neural network. In a supervised neural network, the network relies on training data. In an unsupervised neural network, there is no training data provided, the network tries to creates the correlations on its own and uses these to classify new data. In a reinforcement neural network, the network learns by means of penalties and rewards for right and wrong decisions. [7]

A basic three-layer network can be seen on the left side of the diagram in figure 4.

2.4. Convolutional Neural Networks

A subset of neural networks that has gained much attention in visual recognition is the convolutional neural network. The structure of CNNs allows for the learning of abstract feature detectors and allows mapping of these features into representations. These representations enhance performance of future classifiers [8].

CNNs have an architecture with multiple stages that each contain three layers. These layers include a convolutional layer, a pooling layer and an output layer (fully-connected layer). The convolutional layer is where the primary processing is done by having a spatially small filter slide, or convolve, over the full volume. This in turn produces an activation map that is two dimensional. The pooling layers main functions are to reduce the spatial size, number of parameters and control overfitting. The output or fully-connected layer contains the final scores of the classification [6]. CNNs have become specifically useful with RS data for scene understanding, target recognition and pixel classification [8].

Convolutional networks deal with tensors, which are in essence nested arrays. The layers of CNNs are arranged in three dimensions that also include depth. Images are defined as three dimensional objects which include colour encoding. In figure 4, is a CNN (right side) compared to a regular three-layer neural network (left side). The image input in the CNN is shown as the red block with the dimensions of the image being the height and width. The Red-Green-Blue (RGB) channels make up the depth.

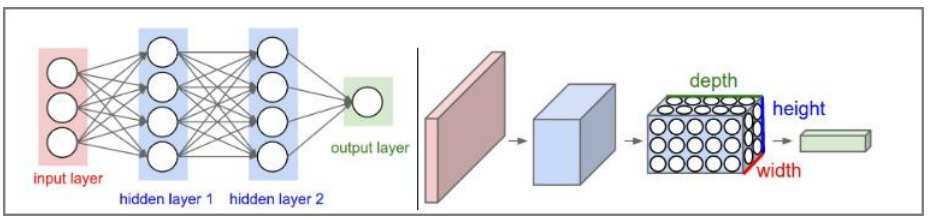

Figure 4: Comparative diagram of a three-layer neural network (left) vs a CNN (right) [6]

3. CHALLENGES INHERENT TO EARTH OBSERVATION AND REMOTE SENSING

According to Kanevski et al. [9], geospatial data in general, has very specific characteristics or 'particularities' that complicate analysis and prevent modelling via traditional geostatistical models. Amongst these are nonlinearity, spatial and temporal non-stationarity, multi-scale variability, presence of noise and extremes/outliers, and a multivariate nature [9].

Additionally, Ball, Anderson, and Chan [10] after reviewing 57 survey papers in DL and RS and 205 RS applications papers compiled a list of nine challenges for DL in RS. The issues they identified are below in table 1.

One of the issues listed below by Ball, Anderson, and Chan [10], is a "high barrier to entry" as a challenge for the RS community. Until a few years ago, most ML tools were made by software developers for software developers. These tools also use a variety of programming languages and proficiency in the language is typically a requirement to use them. It is only in the last few years that there has been a focus on making tools easier to use for non-developers.

The size of the data is also restrictive for most independent or student researchers to process on personal computers as processing generally requires systems with multiple graphic processing units (GPUs).

Table 1. Challenges for deep learning in remote sensing [10]

	Challenge	**Details**
1	Limited data sets/limited training data	In all the papers surveyed, there were five commonly used data sets. They showed that overall accuracies can not necessarily be trusted based on the number of training samples for each paper. They also showed that the data sets are saturated. They recommend that new data sets are required. In RS, there is only a small set of imagery with samples labelled for training.
2	Models for RS applications are often very complicated	RS models can have very intricate relationships and can be inaccurate if the input data does not take this into account. Low temporal resolution can be a challenge as well. The recommendation is to focus on more complex features instead of pixel-level and spatial patterns
3	Big data	Algorithms need to be streamlined and there needs to be better processing power. There is a focus on being able to combine different types of data.
4	Non-traditional data sources	Using sources such as social media photos and videos or tweets with geo-location data for real-time analysis.
5	DL architectures	Complex RS problems may not be solvable with current DL architectures.
6	Transfer learning	Current challenges include, transfer when endmembers are not the same, transfer of low to midlevel features, especially those from different domains, transfers for imagery collected in different atmospheric conditions and times.
7	An improved understanding of DL systems	New DL methods, both practical and theoretical need to be explored to go deep. There should be improved training and generalization capabilities.
8	High barriers to entry	Hardware restrictions and multiple software development requirements can create a steep learning curve for DL. Many RS tasks are not included in standard libraries.
9	Training and optimizing the DL	There are many ways to train a DL system and it can be difficult. DL systems can also have millions of parameters.

Below in table 2, is a summary of some of the RS applications from twenty additional papers that were surveyed by Ball, Anderson, and Chan [11]. This paper's aim was to, "showcase what has been done, what is being done, and what big questions remain and need to be tackled by the community." [11]. They identified CNNs as being one of the primary algorithms used for remote sensing data, often deep neural networks. They also found it was common to use non-remote sensing pre-trained data as well as transfer data to assist in classification.

Table 2. Challenges and contributions in RS applications [11]

RS Application	Challenges	Example Contributions
Synthetic Aperture Radar (SAR) processing	Traditional SAR processing methods use features crafted by hand	• CNNs for feature extraction allow for change detection and classification • Algorithms that require no prior processing or segmentation
Ocean processing	Ships are very small, cloud interference, wave interference	• Deep CNN to extract features and detect ships • Provided bounding boxes for recognition of ocean fronts
Classification and labelling	Large size of imagery, multiple resolutions, image matching	• CNNs for the recognition of dust, smoke, hurricanes, etc. • Deep CNNs for detection buildings from orthoimages
Multimodal (mixed techniques)	Combining multiple technologies	• CNN detector for golf courses, augmented with temporal data • Hyperspectral and visible images combined with CNN for feature extraction later also combined with spectral, statistical and spatial data
Spectral-spatial processing	Anomalies in hyperspectral data	• Stacked denoising autoencoder for hyperspectral anomaly detection • Deep stacked sparse autoencoder for feature learning in hyperspectral images
Object tracking and recognition	Large spatial areas, drifting in long term tracking	• Dual correlative deep networks for aircraft recognition • CNN spatial clustering and chip detection for the identification of surface-to-air missile sites
Architectural studies	Determining if shallow CNNs are sufficient for feature recognition	• For remote scenes, greater CNN depth is essential for identifying features • Over time, CNNs have become deeper

3.1. Training data

The size of the data is also a major factor in being able to train a system effectively. Ball, Anderson, and Chan [10], identified the challenges associated with training for DL systems which deal with vast amounts of RS data.

The training issues they identified include:

- DL systems can have millions of parameters
- RS data may not be labelled
- Hyperspectral data is a very large data cube with many layers, while DL algorithms are typically trained from very small RGB images
- Light detection and light detection and ranging (LiDAR) have insufficient literature as the data is not an image, but a point cloud
- Gridded searches or random methods are required for optimization which can be very time consuming [10]

Helber et al. [12], have also emphasized the importance of having a high-quality dataset for training and classification. They state that one of the challenges to creating these training sets has been access to ground truth datasets that are reliably labelled. In response to this challenge, they created a dataset called EuroSAT from Sentinel-2 images. Their dataset consists of 27,000 images consisting of 10 different classes which can be seen in figure 12. For instance, in the river, and sea and lake classes, they have tried to accommodate the various colours, locations and bodies of water. Each of the 10 classes contain 2000 to 3000 images. Each image measures 64x64 pixels and is at a 10 m per pixel resolution.

Approximately 1.6TB of data in the form of compressed images comes from the Sentinel-2 constellation every day. Each of these 27,000 images had to be manually checked and sorted multiple times in order to get an acceptable accuracy for their training set [12].

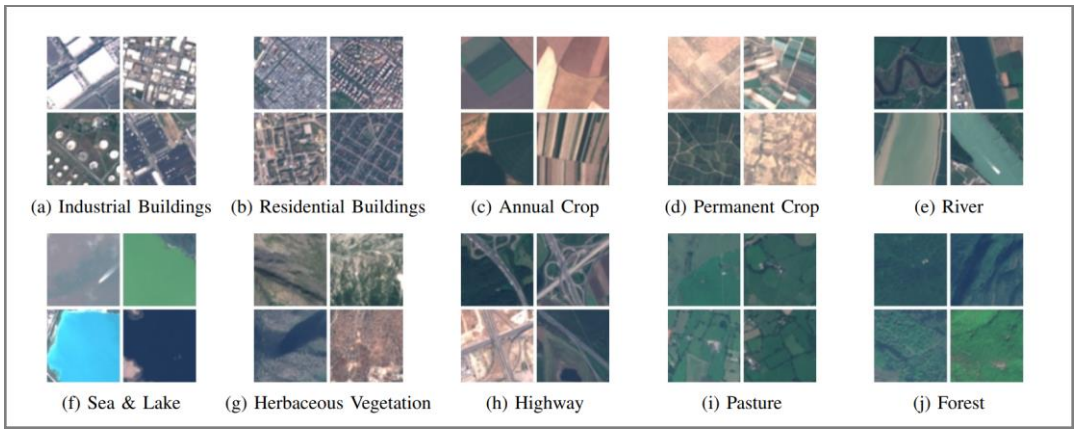

Figure 5: EuroSat training dataset from Sentinel-2 images [12]

4. REAL WORLD EXAMPLE: DISCOVERING ALGAL BLOOMS WITH DATA MINING AND MACHINE LEARNING IN EARTH OBSERVATION

4.1. Phytoplankton, cyanobacterial and algal blooms, eutrophication

The photo below in figure 6, which is of Gotland, a Swedish island in the Baltic Sea, is called 'Van Gogh from Space'. It is a part of the United States Geological Survey (USGS) 'Earth as Art' image gallery [13]. Aside from being a beautiful image, the image also has the significance of showing a 'phytoplankton bloom'. This image was taken on the 13th of July, 2005 from Landsat-7. The phytoplankton biomass is caused by an increase in nutrients, generally associated with rising nitrogen concentrations. These occurrences are often called algal blooms and are also associated with cyanobacteria blooms, which have the potential to be toxic or harmful. They can be natural and seasonal or caused by pollution originating from densely populated areas or industrial runoffs [14]. The red arrows in figure 6 show these algal blooms. The image below is a colour enhanced image and the blooms are seen in bright green. Landsat-7 bands and the appearance of colours in feature properties will be discussed in section 4.2.

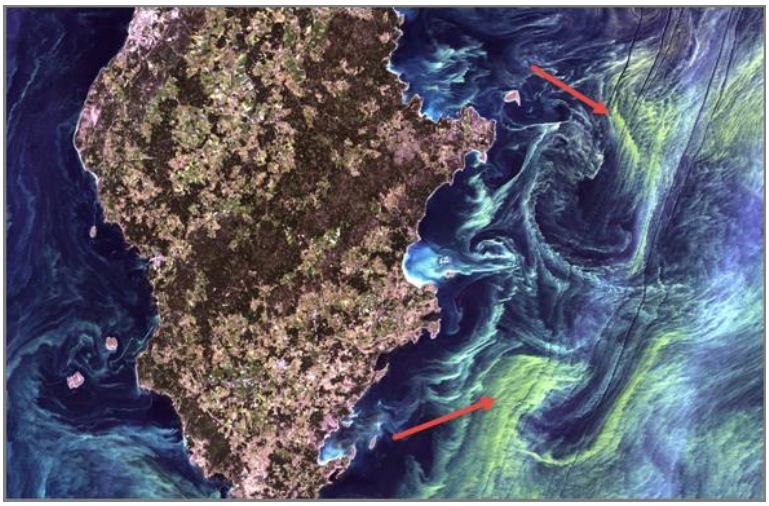

Figure 6: Gotland Island on 13 July, 2015. Red arrows indicate algal blooms [13]

A bloom occurring because of related pollution is known as eutrophication. The blooms are also correlated with an increase in chlorophyll a which contributes to the green color of the blooms [14]. The seasonal variations of these blooms in the Baltic Sea are seen generally during July and August when the water is warmer [15]. In the Gotland Sea, these seasonal blooms have also been known to occur in the autumn months from October through December [14]. Signs of blooms outside the seasonal windows can be possible indications of eutrophication.

Up until the 1960s, most blooms which were the result of eutrophication were recorded only in coastal waters. After the 1960's it became common to see these blooms occurring in the open areas of the Baltic Sea [16].

4.2. Tools and methodology: Google Earth Engine

The Google Earth Engine has been chosen as the tool for this study. GEE currently is able to display 40 years of historical satellite imagery, including publicly available Landsat-7 RS data catalogues, and has a built in JavaScript API that makes geospatial analysis across petabytes of data possible in the Google Cloud Platform [17]. This means that a local installation is not necessary and all computations can be executed in the cloud. As these are extremely large datasets, this is one of the major advantages and reasons for choosing GEE. GEE also has a Python API that is equipped with Cloud Datalab which allows it to be run with Jupyter notebooks. The JavaScript API is available in the GEE coding editor and is an online Integrated Development Environment (IDE) which allows for immediate visualization and rapid prototyping [17]. The IDE can be seen in figure 7.

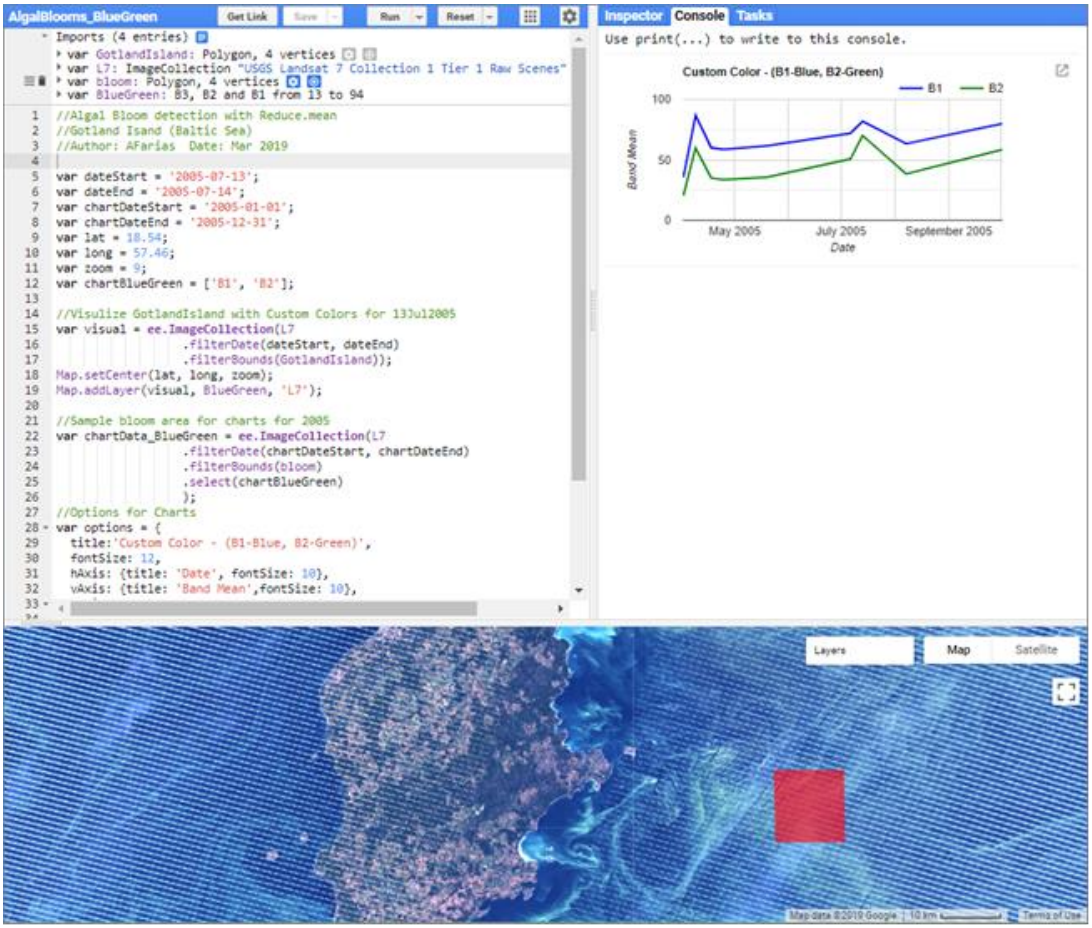

Figure 7: Earth Engine IDE. Shown is the JavaScript editor (top left), console with a chart in the console output (top right), and map visualization layer (bottom)

Earth Engine, although relatively new, has already seen hundreds of scientific papers published making use of the tool for a variety of applications such as medical studies, vegetation and forestry, wetlands and hydrology, agriculture, urban studies and disaster management [18].

4.3. Landsat-7 data catalogue in Earth Engine

Landsat-7 was launched on the 15th of April, 1999 and is currently still operating today. At the time of submission of this report, Landsat-7 will be celebrating its 20th year anniversary in orbit, but is expected to be replaced by Landsat-9 late 2020 [19]. Landsat-7's system capabilities include high volume, high resolution and multispectral resolution while averaging 250 scenes per day. It was designed for a 705 km, sun synchronous orbit with a 16-day mapping cycle. It also has an internal cloud cover prediction mechanism and only captures sunlit areas which prevent it from collecting data that is unusable [20].

ETM+ has a swath of 185 km, with six spectral bands (1-5, 7), a panchromatic band (8) and a thermal band (6). It has a spatial resolution of 30 m for the spectral bands, the panchromatic band has a resolution of 15 m and the thermal band has a 60 m resolution. All bands have two gain settings of high or low [19],[20].

4.3.1. Earth Engine's implementation of Landsat-7 data

In |GEE, the Landsat-7 data catalogue, which has been acquired from USGS, has designated 10 bands for use in analysis. This can be seen below in table 3. The bands are selectable in layers of up to three bands to create a composite image in GEE. In GEE, the gain settings are only separated in thermal band 6 as B6_VCID_1 and B6_VCID_2. Band 6 has also been resampled from the original 60 m resolution to 30 m.

The dataset used for this study is the USGS Landsat 7 Collection 1 Tier 1 Raw Scenes collection (LANDSAT/LE07/C01/T1). Tier 1 describes scenes with the highest available data quality. These scenes are appropriate for time-series analysis as they have been calibrated across the various Landsat sensors and they have Level-1 Precision Terrain (L1TP) processed data [17].

Table 3. Landsat-7 Band Details for GEE [19],[20], [21]

Band Name	Resolution	Wavelength	Description
B1	30 m	0.45 - 0.52 µm	Blue
B2	30 m	0.52 - 0.60 µm	Green
B3	30 m	0.63 - 0.69 µm	Red
B4	30 m	0.77 - 0.90 µm	Near infrared
B5	30 m	1.55 - 1.75 µm	Shortwave infrared 1
B6_VCID_1	Resampled from 60 m to 30 m	10.40 - 12.50 µm	Low-gain Thermal Infrared 1
B6_VCID_2	Resampled from 60 m to 30 m	10.40 - 12.50 µm	High-gain Thermal Infrared 2
B7	30 m	2.08 - 2.35 µm	Shortwave infrared 2
B8	15 m	0.52 - 0.90 µm	Panchromatic
BQA			Landsat Collection 1 QA Bitmask

4.4. Recognizing algal blooms with remote sensing data from Landsat-7

The measures for computing a body of water as eutrophic include secci-disk transparency (SDT), total phosphorus (TP) and chlorophyll-a (Chl-a). Chl-a measurements are not influenced by sediment or acids and correlate with the volume of phytoplankton concentration in a body of water. The increase in Chl-a is a good indicator for detecting blooms specifically with RS data [22].

Fuller, Aichele, and Minnerick [22], determined in Landsat 7 images, that to detect Chl-a, "…the combination of band 2 (Green), band 3 (Red), and band 7 (short wave infrared) produced the highest R2 values." R2 is the coefficient of determination and gives a statement about the error or the residual, as shown in chapter 2.2.3.6, between the predicted and the expected value. The higher the value is the better the prediction is.

Therefore, the recognizing of chlorophyll-a (variable Chl-a) from Landsat-7, for Fuller, Aichele, and Minnerick [22], can be seen in equation 1, where the variables a, b, c and d are the derived coefficients from the regression equation. For the purposes of identifying this configuration of bands, it has been labelled it as 'FAMChl'.

$$\ln(Chl - a) = a(band2) + b(band3) + c(band7) + d \quad (1)$$

Weber [23], has specified that in order to identify cyanobacteria specifically, and not green algae, a 620 µm band is necessary. This falls in between band 2 and band 3 and is not covered by the instrumentation on Landsat-7. Therefore, any detections with Landsat-7 are assumed to only be algal blooms and make no assumptions about corresponding cyanobacteria levels.

4.5. Visualization of band configurations

A number of other configurations were looked at for band optimization. Images from the visualization of the following band composites can be seen in figure 8. These include the following:

- RGB 'True Color' – (B3,B2,B1)
- False Color – (B4,B3,B2)
- Short-Wave Infrared (SWIR) – (B7,B4,B2)
- FAMChl – (B3,B2,B7)
- Custom Color – (B3, B4, B7)

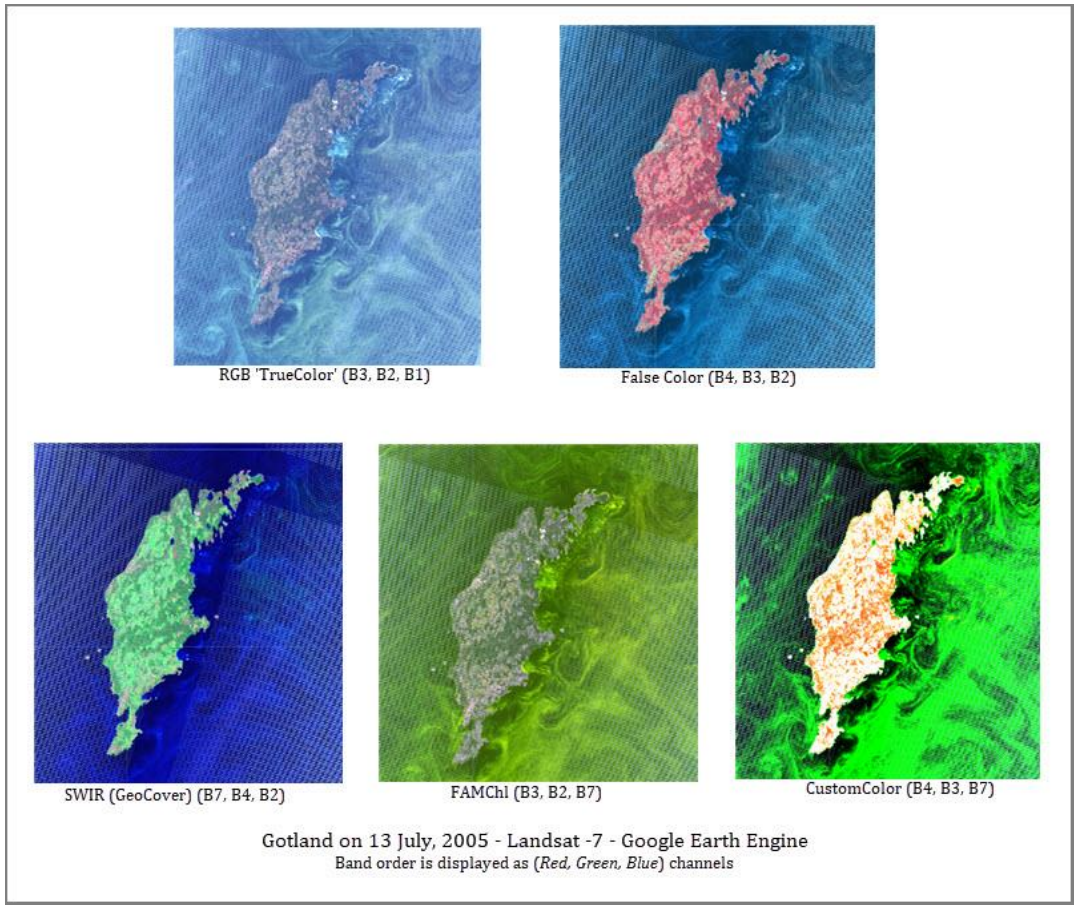

Figure 8: Visualizations of band combinations tested for detecting algal blooms

The artefact lines that can be seen in the images figure 9 are part of a Scan Line Corrector (SLC) failure that happened on Landsat-7 on May 31st, 2003. Despite the SLC fault, a USGS report found that the data was still excellent quality for at least 86% of the pixels when augmented with interpolation [24]. All Landsat-7 images have been processed with the SLC in 'off' mode since the fault.

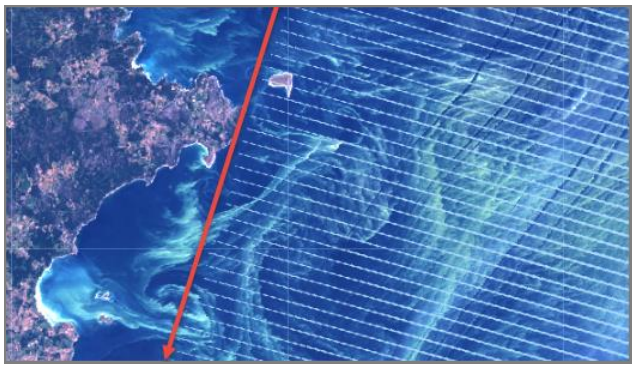

Figure 9: Visualizations of band combinations tested for detecting algal blooms

The colour of primary features such as vegetation or water changes significantly depending on the combination of bands chosen. This is visually significant, but also significant when sampling data at the pixel level for scene analysis. Feature colours can be seen in the table in figure 10 for true colour, false colour and SWIR.

	True Color **Red: Band 3** **Green: Band 2** **Blue: Band 1**	**False Color** **Red: Band 4** **Green: Band 3** **Blue: Band 2**	**SWIR (GeoCover)** **Red: Band 7** **Green: Band 4** **Blue: Band 2**
Trees and bushes	Olive Green	Red	Shades of green
Crops	Medium to light green	Pink to red	Shades of green
Wetland Vegetation	Dark green to black	Dark red	Shades of green
Water	Shades of blue and green	Shades of blue	Black to dark blue
Urban areas	White to light blue	Blue to gray	Lavender
Bare soil	White to light gray	Blue to gray	Magenta, Lavender, or pale pink

Figure 10: The colour of feature display in composite images [25]

4.6. Historical remote sensing and ground truth comparisons

Since 2002, the Swedish Meteorological and Hydrological Institute (SMHI) has been monitoring algae levels in the Baltic Sea. They are now, since 2009, supplementing traditional water sampling methods with satellite data from ENVIronment SATellite (ENVISAT) and Earth Observing System Aqua (EOS-AQUA) using the MEdium Resolution Imaging Spectrometer (MERIS) and Moderate Resolution Imaging Spectroradiometer (MODIS) sensors respectively [26]. The sensors are only able to detect surface level algae which have a high reflectance. They are not able to see algal blooms currently through clouds or at night [26].
Hansson and Hakansson [27] described the 'Baltic Algae Watch System' which monitored cyanobacterial blooms from 1997-2006. They processed images from the National Oceanic and Atmospheric Administration (NOAA) - Advanced Very High-Resolution Radiometer (AVHRR) (NOAA-AVHHR) based on a supervised classification algorithm in near infrared and thermal channels. The NOAA-AVHHR data, however, has a poor resolution at ~1 km^2 which made costal detection difficult [27].

In 2005, there were 13 Cholorphyll-a measurements taken at the 'BY15 GOTLANDSDJ' (Gotland Station). The station is marked by the red x in figure 11. This number of samples varies from year to year. For instance, there were 11 samples taken in 2016 [28].

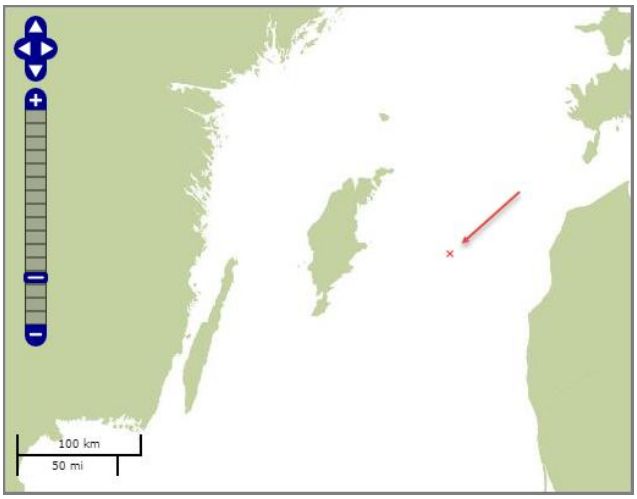

Figure 11: Location of SMHI Chlorophyll-a sampling (SMHI, 2019b)

The data for these measurements is publicly available on the SMHI Swedish Sea Archives/Svenskt HavsARKiv (SHARK). The SHARKweb has data collected by SMHI that goes back to 1893 for numerous marine biological, chemical, and physical parameters [29].
The 13 samples for Chl-a that were taken in 2005 can be seen in figure 12. It should be noted that the dates of sampling do not necessarily correspond with the dates of blooms or the dates that Landsat-7 was acquiring data.

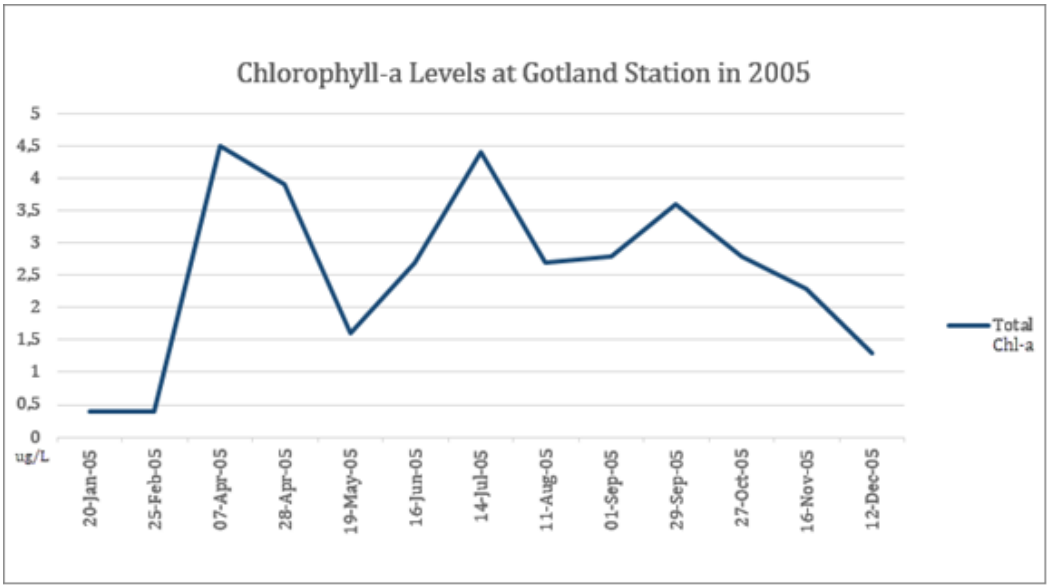

Figure 12: Chlorophyll-a levels sampled from Gotland station in 2005, from the SMHI database for use in the ground-truth comparison

4.7. Mean reduction by geometry region in Earth Engine

GEE has the ability to define specific geometry objects to areas on the map that are to be used for analysis. For the purposes of this study, a rectangle around Gotland Island was defined. This can be seen in light blue in figure 20. The darker blue around it is the result of returned Landsat-7 data for that day. Multiple swaths can be in the image depending on the angle that the image was acquired. In figure 20, there is only one swath of 185 km in width. The area of the Gotland Island geometry is ~32,000 km, the area of the bloom geometry is ~100 km.
The sample size of ocean was approximately 10km^2 as seen in red in figure 13. A test was also done by increasing the number of sample blocks to three, however the accuracy decreased significantly. All further tests were done using one sample block.

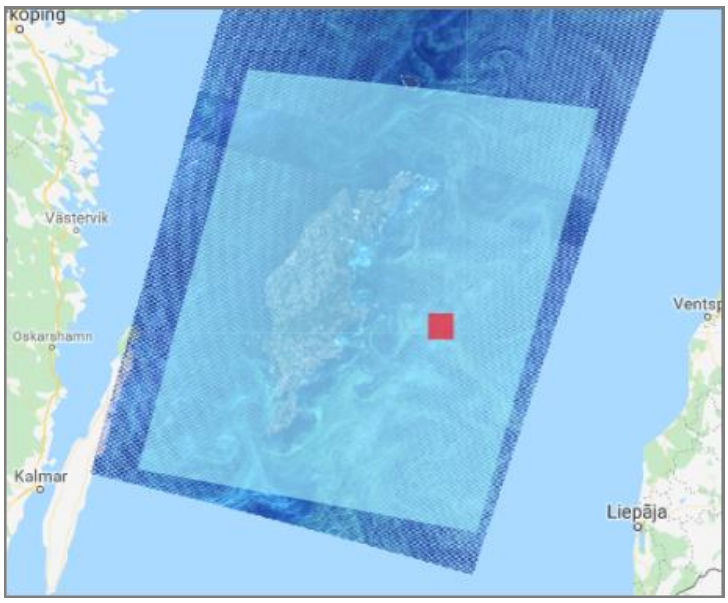

Figure 13: Geometry objects used for analysis in GEE. (Light blue) Gotland Island (Red) bloom sampling location

The ee.Reducer is a way to get pixel statistics of a geometry or region. The reducer will take an area in the image and compute a value for each of the bands. If it is a true colour image with RGB bands, it will return three numbers, one for each band. In each of the band combinations above, there are three bands, so what has been done is to take the mean of the pixels and between the bands. All bands must be specified to be sampled at the same resolution, in this case, the sampling was done at 30 m as the bands all have this spatial resolution. This is especially useful when comparing spectral values over a time series. It is worth noting that the maximum pixels that the reducer can compute is 10 million, so large areas, such as the whole island cannot be computed. This will work with smaller areas such as the bloom area. There is also a function to only use the maximum number of pixels. The result from this will be a random selection of pixels up to 10 million. A diagram of how ee.Reducer is used with the workflow can be seen in figure 14.

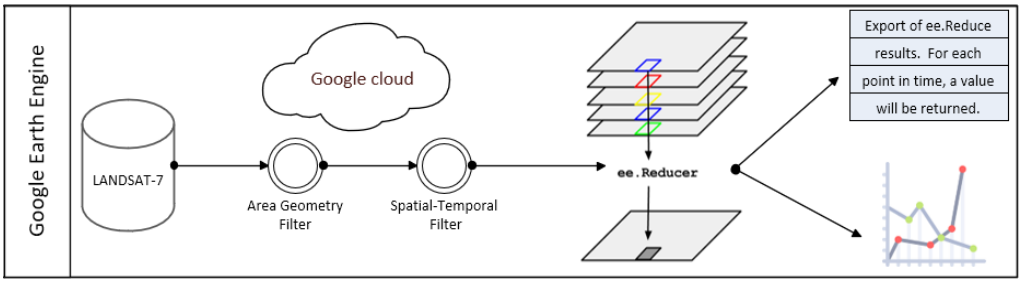

Figure 14: Workflow of ee.Reducer being applied to Landsat-7 with an area geometry filter and a spatial-temporal filter. [29]

The band mean spectral values for CustomColor, TrueColor, FalseColor, FAMChl and SWIR were calculated by using ee.Reducer.mean. The values for Chl-a levels as well as the band mean spectral value were plotted seen in figure 15. The results of this exercise also showed that the sampling dates for Chl-a and the collection date for the Landsat-7 images were non concurrent. There were 13 sample dates for Chl-a and 9 dates for Landsat-7 data. Of these, only one date perfectly matched. In the end, only 19 sample dates were used as there was no Landsat-7 data for Jan, Nov or Dec. There are many reasons for why this data was missing. Landsat-7 is meant to have a 16-day return cycle, but it is also meant to discard any scenes with cloud cover or any night time scenes. It is also possible for scenes to be discarded where there are system faults or calibration errors.

In the case of algal blooms in the Baltic Sea, they are often seasonal and can last an extended period of time, so an assumption was made that the last sampling or collection value would remain until the next one was received. This could lead to an uncertainty for a detailed analysis of time periods with an offset to a certain data point. For instance, it would be incorrect if a Chl-a sample was taken when there were no blooms and a bloom only appeared a few days after when the satellite passed the region. This method did, however, allow for a rough trend to be seen when plotted and a preferred band was chosen as a result. A more accurate regression model/curve would be possible with more row data and more data points.

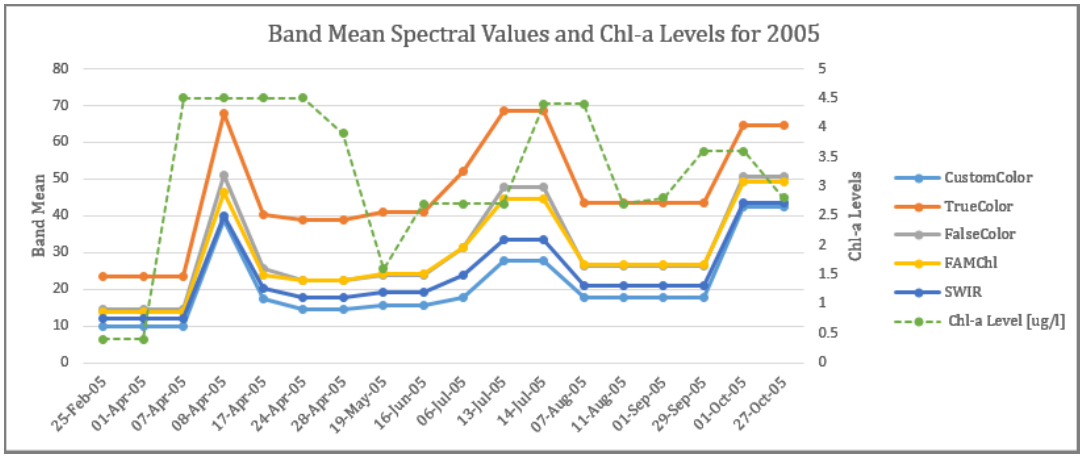

Figure 15: The lack of concurrent sampling points from SMHI and Landsat-7 image data means that data over a larger period is needed

As a backup to the method above, all bands were also plotted individually. By doing this, it was shown that bands 1 and 2 were preferred over band 3. As such, for all future analysis only band 1 (blue) and band 2 (green) were used for analysis. This can be seen in figure 16.

According to Pitarch et al., [31] there are two types of algorithms presently being used to measure Chl-a via remote sensing. This includes an empirical method using a blue/green reflectance ratio, as we have identified above. The other method is a semi-analytical method, where the water type is more complex in the possible colorations found. Case I waters are generally found in open waters where as Case II waters can be anywhere the water is discoloured. They have considered the Baltic Sea to be a "challenging test bed for remote sensing" with a high concentration of coloured dissolved organic matter (CDOM). The standard algorithm that is provided for Chl-a detection, generally by the space agencies, is one that is specific to global applications and does not take into account these discolorations specific to certain regions [31].

4.8. Results: Analysis with blue and green bands

Since Landsat-7 launched in 1999, it was decided to do a ten-year initial analysis through to 2009. After 2009, the SMHI changed the way they observed Chl-a via remote sensing, so a comparative analysis by year is challenging beyond this period. For a more recent analysis, the years of 2010-2019 can be also studied together, however, that is out of scope for this paper. Nevertheless, the results in this paper can be used for further analysis and studies of the earlier period.

The numbers for band 1 and 2 returned from the initial test are shown in figure 16. The chart is a sample of the charts included in the GEE IDE. As seen in this plot, there are a number of spikes above 100. These were cross checked to bloom dates with SMHI and did not correspond. Also, in all the tests with the 2005 data, there were never any points above 90.

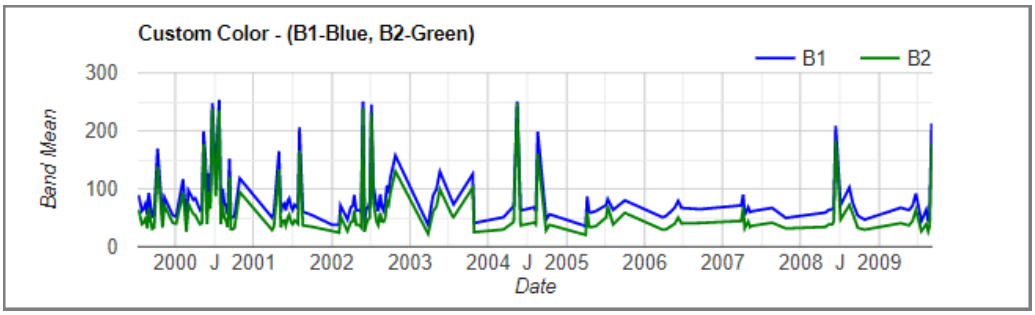

Figure 16: Landsat-7 band 1 and 2 spectral values for the bloom geometry with abnormal spikes above 100 (1999-2009)

The individual spectral values were exported and evaluated against visual representations of the plotted days in question. Anything over 100 was considered an anomaly and in most cases, when cross checked visually, was due to cloud cover or what appeared to be instrument errors. There was also a very basic preliminary comparison of the number of days of known cyanobacterial blooms for the period of 1999-2009. For this, the spectral values have been compared with heat maps from SMHI with the number of days with cyanobacterial blooms [26].

The next evaluation that was done was to attempt to match days with data points for both Chl-a sampling and satellite imagery. For the period of 1999-2009, there were 137 days with Chl-a data samples. For Landsat-7, there were 145 days. These were matched with an interval of +/-7 days. This left 96 days over the ten-year period. Because there were some unexplained discrepancies in data around 2006, possibly due to recalibration, it was decided to focus on the years of 2001-2005 (minus SLC fault data). The final analysis contained 45 days and can be seen in figure 27.

The SCL fault occurred on the 31st of May, 2003 and data from the satellite became temporarily unavailable. The data became available again in July, but a second product with the gaps filled in via interpolation was not available until May 10, 2004 [20]. All data points from this time were also removed as seen on the x-axis in figure 17.

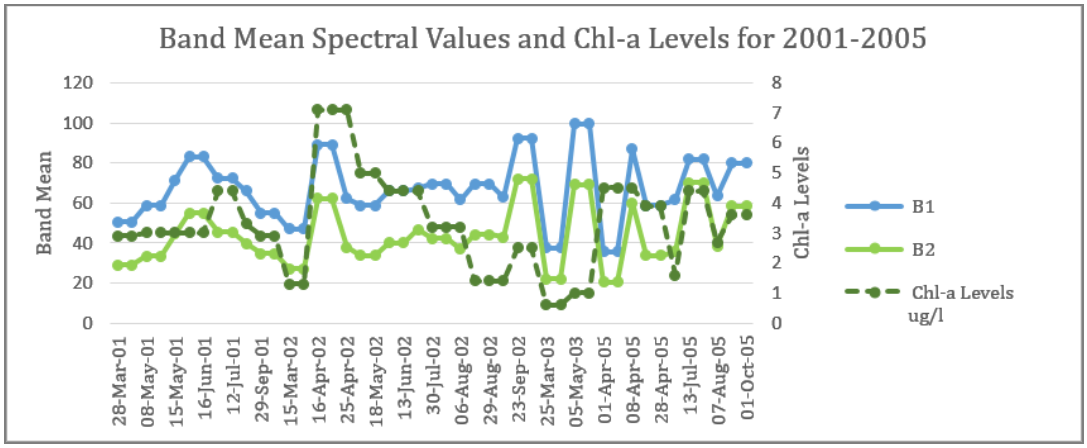

Figure 17: Bands 1 and 2 plotted as spectral values by date shown with Chl-a sampled levels

While there are a number of correlations between the collected samples and spectral values in the blue and green spectrum, more detailed analysis of individual points is necessary. While we have seen that cloud cover should be filtered from the collection, certain outliers were still found, and it is possible that a partially clouded scene can be skewing results. There are a number of cloud filtering algorithms that can be explored in future studies.

5. Discussion and Recommendations for Future Study

The study of detecting algal blooms via earth observation was chosen because it has been identified as having seasonal events that are regularly monitored via satellite. As an added benefit, it is also monitored via ground truth methods such as water sampling. This made it possible to test various detection techniques and cross check them with historical results. Gotland Island was chosen as there was a very large bloom that extended over a longer period in 2005. The algal blooms in this area also have a very high reflectance which can be picked up with remote sensing.

There were a number of discoveries in the analysis of the earth observation data. One of the most important ones is that ground sample dates do not necessarily correspond with Landsat-7 pass over dates. There is also a chance that the data collected is not useable, such as with heavy cloud cover. Because of this, it is recommended that a future study be done with a satellite or a constellation of satellites that have a more frequent revisit time. Also, in order to study cyanobacterial blooms as well as algal blooms, it is recommended to use a 620 µm band which Landsat-7 does not have [22].

The area that was sampled for this study was ~10 km^2, which is a very large area. There was also only one sample location as the initial tests with multiple locations showed a decrease in accuracy. This should be further studied with various sample areas and sample methods. Also, other sample methods might be necessary for different bodies of water with other algal types. The tested areas should also include coastal areas which can have vastly different colorations depending on runoffs and surrounding terrain.
This exploration was done specifically using the result of single spectral value which was calculated for a sample location. There are also other random sampling methods that could be

used and tested for better results. The algorithms that have been previously packaged in ocean observation software are fairly complex and have the need for customization to specific areas [31]. One future potential that has arrived from this study could be building a training set of historical algal blooms from the data points identified above.

The GEE JavaScript API was excellent for fast prototyping and testing ideas, however, for more powerful algorithms, extended libraries should be considered for in depth studies. The Python API is a good candidate for this, and Google is also working on a number of solutions for unsupervised learning based on the Weka platform [31].

It is worth mentioning the computer processing time involved in preforming the analysis over multiple years. In previously used Geographic Information Systems (GIS), such as ArcGIS, this type of analysis would have taken hours or even days [23]. Because it is hosted in the Google cloud, the processing for ten years of data took a matter of seconds. This is the biggest take away from this exploratory study with Google Earth Engine. The ability to rapidly prototype and visualise results across vast datasets, within seconds, has the potential to dramatically change earth observation data mining techniques.

6. CONCLUSION

The areas being studied, data mining and machine learning, are wide-ranging fields of study. In EO, there is an increasing focus on machine learning to handle the massive amounts of data that are being collected on a daily basis by higher resolution instruments. There are many challenges in EO, including the size of data, its variable and complex nature, a high barrier to entry, and the datasets used for training the data. However, as the field grows these challenges are being addressed.

Software companies are attempting to combat the barrier to entry by providing easy to use IDEs which enable fast prototyping and almost instant visualizations. A new training data set which is focused on improving training for deep learning systems, EuroSAT, has been implemented
In this paper as a real-world example, the island of Gotland in the Baltic sea was studied by using GEE. Ten years of Landsat-7 data, along with a dataset from SMHI, for ground truth, was analysed to discover historical algal blooms in the Baltic Sea.

One of the biggest take-aways from this study is the speed of processing in GEE, which is hosted in the Google cloud. Where previously, this type of analysis would have taken hours or days to process, it is now available in a few seconds. While GEE is still very new, there have already been hundreds of scientific papers published using it.

Although primarily an exploratory study, this paper has shown the increasing potential for new tools and techniques to enhance the analysis of earth observation data for scientific research.

REFERENCES

[1] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, “Google Earth Engine: Planetary-scale geospatial analysis for everyone,” Remote Sens. Environ., 2017.
[2] P. Berkhin, Survey of Clustering Data Mining Techniques. 2002.
[3] Oracle, “Oracle® Data Mining Concepts 11g Release 1 (11.1),” 2008.
[4] J. Han, M. Kamber, and J. Pei, “Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems),” 2011.
[5] Skymind Inc., “A Beginner’s Guide to Neural Networks and Deep Learning | Skymind,” 2019. [Online]. Available: https://skymind.ai/wiki/neural-network. [Accessed: 29-Mar-2019].

[6] A. Karpathy, "Convolutional Neural Networks for Visual Recognition," 2018. [Online]. Available: https://cs231n.github.io/. [Accessed: 29-Mar-2019].
[7] J. Alzubi, A. Nayyar, and A. Kumar, "Machine Learning from Theory to Algorithms: An Overview," in Journal of Physics: Conference Series, 2018, vol. 1142, no. 1.
[8] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," IEEE Geosci. Remote Sens. Mag., vol. 4, no. 2, pp. 22–40, 2016.
[9] M. Kanevski, A. Pozdnukhov, V. Timonin, and A. Pozdnoukhov, "Machine Learning Algorithms for GeoSpatial Data. Applications and Software Tools," 2008, vol. 12.
[10] J. E. Ball, D. T. Anderson, and C. S. Chan, "Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community," J. Appl. Remote Sens., vol. 11, no. 04, p. 1, Sep. 2017.
[11] J. E. Ball, D. T. Anderson, and C. S. Chan, "Special Section Guest Editorial: Feature and Deep Learning in Remote Sensing Applications," J. Appl. Remote Sens., vol. 11, no. 04, p. 1, Jan. 2018.
[12] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Introducing Eurosat: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification," in IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, 2018.
[13] USGS, 2018. Earth Resources Observation and Science (EROS) Center. [Online] Available: https://eros.usgs.gov/image-gallery/earth-art-3/van-gogh-space [Accessed 22 Nov. 2018].
[14] N. Wasmund and S. Uhlig, "Phytoplankton trends in the Baltic Sea," ICES J. Mar. Sci., vol. 60, no. 2, pp. 177–186, Apr. 2003.
[15] S. Anttila et al., "A novel earth observation based ecological indicator for cyanobacterial blooms," Int. J. Appl. Earth Obs. Geoinf., vol. 64, pp. 145–155, Feb. 2018.
[16] T. Finni, K. Kononen, R. Olsonen, and K. Wallström, "The History of Cyanobacterial Blooms in the Baltic Sea," AMBIO A J. Hum. Environ., vol. 30, no. 4, pp. 172–178, Aug. 2001.
[17] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," Remote Sens. Environ., 2017.
[18] L. Kumar and O. Mutanga, "Google Earth Engine Applications Since Inception: Usage, Trends, and Potential," Remote Sens., vol. 10, no. 10, p. 1509, Sep. 2018.
[19] U.S. Department of the Interior | U.S. Geological Survey, "What are the band designations for the Landsat satellites?," USGS, 2014. [Online]. Available: https://www.usgs.gov/faqs/what-are-band-designations-landsat-satellites-0?qt-news_science_products=7#qt-news_science_products. [Accessed: 20-Nov-2018].
[20] NASA, "Landsat 7 Science Data Users Handbook," 2011.
[21] Google Developers, "Landsat Collections in Earth Engine | Earth Engine Data Catalog | Google Developers," 2019. [Online]. Available: https://developers.google.com/earth-engine/datasets/catalog/landsat/. [Accessed: 01-Mar-2019].
[22] L. M. Fuller, S. S. Aichele, and R. J. Minnerick, "Predicting Water Quality by Relating Secchi-Disk Transparency and Chlorophyll a Measurements to Satellite Imagery for Michigan Inland Lakes, August 2002. Scientific Investigations Report 2004-5086," USGS. 2007.
[23] S. J. Weber, "Utilizing Geospatial Cloud Computing and Data Analytics for Cyanobacteria Harmful Algal Bloom Risk Mapping in Georgia Piedmont Waterbodies," UGA. 2017.
[24] S. Andrefouet et al., "Preliminary Assessment of the Value of Landsat-7 ETM+ Data Following Scan Line Corrector Malfunction," US Geol. Surv. EROS Data Cent. Sioux Falls, SD, USA, 2003.
[25] El Centro Informático Científico de Andalucía (CICA), "An Introductory Landsat Tutorial," 2019. [Online]. Available: https://huespedes.cica.es/geo/agr/Landsat Tutorial-V1.html. [Accessed: 01-Mar-2019].
[26] SMHI, "Monitoring algae from satellite | SMHI," 2010. [Online]. Available: https://www.smhi.se/en/theme/monitoring-algae-from-satellite-1.11923. [Accessed: 01-Apr-2019].
[27] M. Hansson and B. Hakansson, "The Baltic Algae Watch System - a remote sensing application for monitoring cyanobacterial blooms in the Baltic Sea," J. Appl. Remote Sens., vol. 1, no. 1, p. 011507, Dec. 2007.
[28] L. Wesslander, K; Viktorsson, "Summary of the Swedish National Marine Monitoring 2016-Hydrography, nutrients and phytoplankton," Rep. Oceanogr., vol. 60, p. 92, 2017.
[29] SMHI, "Marine Environment Data | SMHI," 2019. [Online]. Available: http://www.smhi.se/klimatdata/oceanografi/havsmiljodata. [Accessed: 25-Mar-2019].
[30] Google Developers, "Statistics of an Image Region | Google Earth Engine API | Google Developers," 2019. [Online]. Available: https://developers.google.com/earth-engine/reducers_reduce_region. [Accessed: 03-Apr-2019].

[31] J. Pitarch, G. Volpe, S. Colella, H. Krasemann, and R. Santoleri, “Remote sensing of chlorophyll in the Baltic Sea at basin scale from 1997 to 2012 using merged multi-sensor data,” Ocean Sci, vol. 12, pp. 379–389, 2016.
[32] Google Developers, “Unsupervised Classification (clustering) | Google Earth Engine API | Google Developers,” 2019. [Online]. Available: https://developers.google.com/earth-engine/clustering. [Accessed: 01-Apr-2019].

AUTHORS

Alexandria Dominque Farias is currently an IT Application Specialist at OHB InfoSys in Bremen, Germany. She is originally from El Paso, Texas, but spent 12 years living in Cape Town, South Africa. She holds an MSc in Space Studies from the International Space University in Strasbourg, France and Masters in Information Technology from the University of Cape Town. She did her undergraduate work at the University of New Mexico and Flinders University of South Australia. She was an IT Officer for Crystal Cruises and held various software development roles at Allan Gray Pty in Cape Town. She started her career at InfoGenesis in Santa Barbara, California where she worked as a Systems Engineer.

Mr. Sun is a professor of Space System Engineering at the International Space University (ISU) located at Strasbourg, France.

Mr. Sun held several senior executive positions both in China and Europe. He started his career as a system engineer of launch vehicle design and followed as project manager of international satellite launching service in China Academy of Launch Vehicle Technology (CALT). He was a founding member of China Manned Space Agency (CMSA) in 1993 and worked as General Designer Assistant for China Manned Space Program (CMSP) for 8 years, he was in charge of launcher system specification definition, system coordination and interface control among spaceship/launcher/spaceport, and conceptual study for RDV as well. After the maiden flight of Shenzhou spaceship, he moved to Munich and worked as Managing Director of EurasSpace GmbH to promote the Sino-European cooperation in space for 8 years. He founded CASC (China Aerospace Science and Technology Corporation) European Office based in Paris in 2010 and served as premier Chief Presentative in the office for 7 years. He was deeply involved in most of the joint space programs between China and Europe in this period.