Curating Multimodal Satellite Imagery for Precision Agriculture Datasets with Google Earth Engine

B S Wijaya[1,*], R Munir[2], N P Utama[2]

[1] Doctoral Program of Electrical Engineering and Informatics, School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Jl. Ganesha No.10, Bandung 40132, Indonesia
[2] School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Jl. Ganesha No.10, Bandung 40132, Indonesia.

*Corresponding author's e-mail: bagussetyawanwijaya@gmail.com

Abstract. In the era of modern agriculture, satellite imagery has been widely used to monitor crops, one of which is paddy. This paper tries to describe the vegetation indices, climate, and soil index features related to paddy plants and curates a collection of satellite imagery on the Google Earth Engine (GEE). This paper reveals how GEE can be used to collect and process multimodal satellite imagery to form a precision agriculture dataset. The objective of this study is to establish a comprehensive precision agriculture dataset by leveraging multimodal satellite imagery to monitor paddy crops. The data collected as a dataset originates from 306 locations in Karawang Regency, Indonesia, during the 2019-2020 period. In the first step, we identify the relevant features essential for paddy crop analysis. Subsequently, we carefully select image collections within GEE based on these features. Afterward, we perform data acquisition and necessary preprocessing through the Google Colab environment. The results showed that satellite imagery from Sentinel-2 outperforms Landsat 8 in terms of spatial and temporal resolution. Apart from that, the generated dataset successfully captures the growth patterns of paddy plants.

1. Introduction

The agricultural sector as the backbone of meeting global food needs has undergone a fundamental transformation in line with advances in information technology and the development of satellite imagery. This era of modern agriculture has brought about significant transformation and provided a solid foundation for innovation in the agricultural sector. This innovation includes a combination of information technology and the use of increasingly accurate satellite imagery. These innovations provide deeper insight into the agricultural environment. This allows agricultural actors to make more precise and effective decisions [1].

In this context, this paper specifically focuses on the field of food crop agriculture, namely wetland and dryland paddy. As a central aspect of meeting global food needs, paddy farming is the focus. Precision agriculture is emerging as a key paradigm in efforts to increase productivity, reduce environmental impact, and address the challenges of climate change [2]. In this practice of precision agriculture, the role of satellite imagery is increasingly important. Satellite imagery provides an

important means of understanding and monitoring plant growth dynamics, as well as the complex interactions between plants and their environment [3].

This paper will review how Google Earth Engine (GEE), as an innovative technology, plays an important role in collecting, managing, and analyzing multimodal satellite imagery. The main goal of using GEE is to form rich and in-depth datasets to support precision agriculture. GEE is a sophisticated platform that allows users to efficiently access, analyze, and combine satellite image data from various sources [4]. This paper will detail the important features relevant to crop farming within the GEE framework, including selecting a series of images from the GEE that match those features. In addition, this paper will also discuss how the integration of spectral data from various satellite imagery sources can strengthen more accurate monitoring of plant growth and more in-depth predictions of crop yields [1].

In addition to its technical merits, this paper has the broader goal of providing a database to support more in-depth precision agricultural analysis. This research has the potential to support the development of more sophisticated analytical models, including more accurate classification of paddy growth and yield prediction, to address global challenges in the agricultural sector [2].

2. Study Area

This study was carried out at Karawang regency (latitude 5º56` to 6º34` S and longitude 107º02` to 107º40` E) in West Java, Indonesia (Figure 1). Karawang district, like other regions in Indonesia, has two seasons, dry and rainy. Monthly precipitation in Karawang ranges from 56.2 to 194.8 mm, surface temperatures range from 16.6 to 32.2 Celsius, and wind speeds range from 0.0 to 8.2 meters per second. Due to the favourable conditions, paddy production in this region reached 1.22 million tons in 2022, covering an area of 204,326 hectares. In this study, satellite imagery of paddy fields was obtained from 306 specific coordinate points during the 2019-2020 period, relying on the Crop Cutting Survey as ground truth. Crop Cutting Survey is a survey conducted by Statistics Indonesia (Badan Pusat Statistik) on a regular basis. The main objective of this survey is to obtain information on the yield per hectare (productivity) of food crops, both paddy and secondary crops. Besides, the survey also collects information on variables affecting productivity, such as cultivation characteristics and government assistances to boost productivity [4].

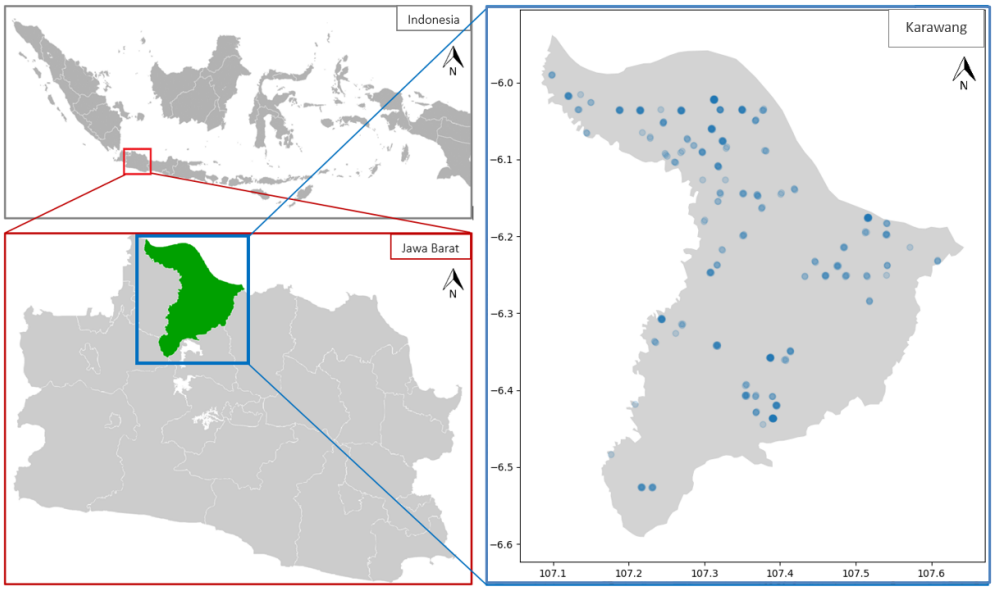

Figure 1. Study area (Karawang, West Java, Indonesia)

3. Theoretical Background

3.1. Satellite Imagery

Satellite imagery plays an important role in realizing precision agriculture by providing visual insight into agricultural land from a height. Satellite imagery is image data taken by satellites orbiting the Earth and includes various visual information about the Earth's surface. This satellite image data is very useful in understanding the dynamics of plant growth, water availability, soil conditions, and environmental interactions in agricultural practices [5]. Satellite imagery consists of various spectral bands that record electromagnetic radiation at various wavelengths, including visible light, near-infrared, and far infrared. Each of these channels provides unique information about the characteristics of the Earth's surface, such as soil conditions and vegetation composition. By analyzing data from these various channels, researchers can understand plant conditions in more depth. In the context of this paper, multimodal satellite imagery, which combines information from various spectral bands and image sources, becomes very important. GEE is an efficient tool for managing, analyzing, and combining multimodal satellite imagery to develop precise agricultural datasets [5].

3.2. Datasets

Datasets in precision agriculture refer to data sets that include a variety of information about agricultural land, including satellite imagery, weather data, soil data, and other relevant information. Appropriate collection and management of datasets play an important role in supporting in-depth analysis. One type of important dataset in precision agriculture is a satellite image dataset. Satellite images from various sources and spectral channels can be used to understand plant growth dynamics, assess plant health conditions, and identify zones on agricultural land that require special treatment [6]. These datasets can cover a certain period, making it possible to monitor changes over long periods. Other datasets include weather data such as precipitation, temperature, and wind speed. The utilization of GEE significantly contributes to data collection and management in the precision agriculture domain. GEE allows users to access and analyse a variety of satellite imagery datasets and geospatial information from various sources on a global scale. Researchers can integrate data from various sources with GEE to develop rich and informative datasets [5].

3.3. Google Earth Engine (GEE)

With its specialized capabilities, GEE is a cloud computing platform uniquely tailored for the retrieval, analysis, and utilization of geospatial data on a highly scalable infrastructure. GEE provides access to an extensive data set of satellite imagery and geospatial information, including data from sources such as Landsat, Sentinel, MODIS, and other satellite data. This platform allows researchers to perform global or local scale analyses quickly and efficiently [5]. In the realm of precision agriculture, GEE assumes a pivotal role in the acquisition, administration, and analysis of multimodal satellite imagery. GEE has parallel processing features that enable fast analysis even on large datasets. GEE users can take advantage of built-in image processing algorithms or develop custom algorithms to perform more specific analyses. One of GEE's advantages is its ability to combine spectral data from various satellite image sources. This allows users to combine information from multiple spectral channels to gain a richer understanding of agricultural conditions. This ability becomes important in understanding the dynamics of plant growth, changes in land cover, and the interaction of plants with their environment. Researchers can perform more in-depth temporal, spatial, and spectral analyses through GEE to develop precision agriculture datasets [5].

4. Methodology

This study is about developing a dataset for data mining, the practical of collecting multimodal satellite imagery data from various image collections. Figure 2 shows the life cycle of a data mining project, as defined as a reference model of the CRISP-DM [7]. CRISP-DM has been used in several studies such as predicting energy consumption [8] and automating seed counts [9]. In the context of a methodology,

it encompasses detailed depictions of the standard project phases along with the associated tasks for each phase. CRISP-DM has six sequential phases, since our goal is to create a dataset, we will only discuss the first two phases, namely business understanding, and data understanding. We will briefly cover the business understanding phase and focus more on data understanding.

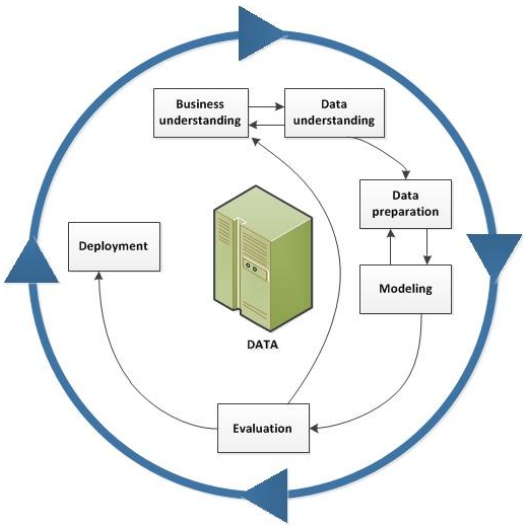

Figure 2. The CRISP-DM life cycle

4.1. Business Understanding

The business understanding phase places its emphasis on comprehending the project's objectives and prerequisites. In this study, we aim to develop a satellite imagery dataset that can be used to carry out analysis related to paddy crops. The conditions needed to do this are data related to crop yields which will be used as field truth data.

4.2. Data Understanding

The data understanding phase entails a comprehensive examination of the available data for the mining process. This pivotal step is instrumental in pre-empting unforeseen issues in the subsequent phase, as it encompasses data access and exploration using tables and visual graphics.

4.2.1. Collecting Initial Data. At this point, we are prepared to access the data and incorporate it into our dataset. Before we start to collect data, there are several things that we must pay attention to, such as which areas we will collect data from, what timeframe we will use, which features from the image collections seem most promising, how to merge various data sources, and how missing values are handled. Potential and tested features that have been used by researchers are inventoried at this stage. Based on these features, we curate the image collection in GEE. Image collections with the best temporal resolution and spatial resolution will be selected for data collection according to the features. Apart from that, the most important thing is that the image collection has data where the research will be carried out because there are several image collections that specifically only provide data for certain regions or countries.

4.2.2. Describing Data. The describing data phase aims to describe and analyse existing data such as data formats, number of records, and field identities, including descriptive statistics and distribution of variables. This helps identify early patterns, trends, and anomalies in the data to understand the data present in the dataset in greater depth. With a good understanding of the data, we can make better decisions about the next steps. A good understanding of the data will help minimize the risk of error and ensure that analysis results are based on quality and relevant data.

4.2.3. Exploring Data. In this data exploring phase, we carry out visual exploration and deeper analysis of the data using visualization techniques and statistical analysis tools. This phase aims to explore new insights about data through deeper exploration and analysis. The main goal is to identify patterns, trends, anomalies, and other valuable information that may not be visible in previous stages.

4.2.4. Verifying Data Quality. The verifying data quality phase focuses on checking and verifying data quality. The main goal is to ensure that the data used is of good quality, free from missing values, outliers, or data errors that can affect model performance and analysis results.

5. Results and Discussion

5.1. Features and Image Collection

Satellite imagery offers valuable features, such as vegetation indices, which can be empirically extracted and modelled to quantify biophysical parameters including plant height, leaf width, and chlorophyll content [10]. The empirical relationship between vegetation indices derived from satellite imagery and in-situ observations for predicting crop yields has been extensively explored in previous research [11]. Plant spectral properties exhibit species-specific variations but share a common underlying pattern. Factors such as plant age, drought stress, or pest infestations can alter the spectral reflectance of leaves. These variations form the fundamental basis for the development of the current vegetation index, essentially represented as a mathematical formula incorporating multiple spectral channels or wavelengths, necessitating its computation using a specific formula [12]. Furthermore, climate and weather data are also frequently employed features. Elevated temperatures, for instance, can negatively impact yields by reducing pollen fertility [13], diminishing the weight of a thousand grains by 4.6%, and decreasing the harvest index by 20% [14]. Exposure to wind and rain can similarly influence rice production by causing paddy plants to lodge [15], thereby affecting grain weight and increasing the proportion of empty grains [16]. Numerous studies have demonstrated that integrating meteorological data with satellite imagery enhances the performance of crop yield prediction models [17] [18]. Table 1 provides an overview of the features employed in various studies for predicting paddy yields.

Table 1. Features used in the precision agriculture domain of paddy rice crop

Location	Image Source	Task	Features	Ref
Bangla-desh	Sentinel-2	Prediction	Vegetation indices (NDVI, NDWI, RGVI, MSI, LAI)	[19]
Nepal	Sentinel-2	Prediction	B2-12, vegetation indices (NDVI), climate (precipitation, temperature, relative humidity), soil	[20]
China	MODIS	Prediction	Vegetation indices (EFI, SIF), climate, soil	[21]
South Korea. North Korea	MODIS	Prediction	Vegetation indices (NDVI, OSAVI, RDVI, MTVI, EVI, LSWI), Climate	[22]
South Korea	MODIS	Prediction	Vegetation indices (NDVI, EVI, LAI, FPAR), climate (precipitation, temperature, solar radiation)	[23]
China	Sentinel-1, Sentinel-2	Prediction	VV, VH, EVI, NDRE, Meteorological data (tmax, tmin, tmean, precipitation, sunshine duration, relative humidity, EAT, AAT, solar radiation)	[24]
China	Sentinel-2	Classification	BSI, LSWI, GCVI, NDVI, EVI, PSRI	[25]
Malaysia	Sentinel-1	Classification	VH	[26]
Malaysia, Indonesia	Sentinel-1	Classification	VH	[27]
Pakistan	Sentinel-2	Classification	B1-12, NDVI, NDWI, NDMI	[28]
USA	USDA-NAIP	Classification	RGB, NIR	[29]

Location	Image Source	Task	Features	Ref
Indonesia	Sentinel-1, Sentinel-2	Classification	VV, VH, B2-12, NDVI, EVI, LSWI	[30]
Brazil	Sentinel-1	Classification	VV, VH	[31]
China	Sentinel-1	Classification	VV, VH	[32]
China	Sentinel-2, MODIS	Classification	Blue, Green, Red, NIR, SWIR1, SWIR2, EVI, WI, Meteorological data	[33]
China	Sentinel-1, Sentinel-2	Classification	VV, VH, B1-12, NDVI, Phenology	[34]
India	Sentinel-1, Sentinel-2	Classification	VV, VH, B1-12, NDVI	[35]
Indonesia	Landsat 8	Classification	B1-7, EVI, NDVI, NDBI, NDWI	[36]
China	Landsat 8	Segmentation	B1-7	[37]
Bangla-desh	WorldView-3	Segmentation	Blue, Green, Red, NIR	[38]

Satellite imagery exhibits variability in spatial resolution (ranging from kilometers to meters), temporal resolution (spanning from monthly to daily), and spectral characteristics. Researchers in the agricultural domain commonly utilize diverse satellite image datasets, including MODIS [21][22][23][33], Sentinel-2 [19][20][24][25][28][30][33][34][35], Sentinel-1 [24][26][27][30][31][32] [34][35], Landsat 8 [36][37]. In some instances, researchers employ multimodal satellite imagery, incorporating Sentinel-2 multispectral data alongside Sentinel-1 radar data [24][30][34][35]. Concurrently, other studies [33] make use of multispectral data from both MODIS and Sentinel-2. A summary of several datasets and image collections available within the GEE is presented in Table 2.

Table 2. Image collections in GEE

Data Provider	Dataset	Image Collection	Band	Resolution	
				Spatial	Temporal
European Union/ESA/Copernicus	Sentinel-2	COPERNICUS/S2_SR_HARMONIZED	B1-12	10-60 m	5-day
	Sentinel-1 SAR GRD	COPERNICUS/S1_GRD	VV, VH	10 m	daily
USGS	USGS Landsat 8	LANDSAT/LC08/C02/T1_L2	SR_B*, ST_B10	30 m	8-day
NASA	MODIS Terra	MODIS/061/MOD09GQ	sur_refl_b01, sur_refl_b02	250 m	daily
	MODIS Leaf Area Index	MODIS/061/MCD15A3H	Fpar, Lai	500 m	4-day
University of California Merced	TerraClimate	IDAHO_EPSCOR/TERRACLIMATE	pr, tmmx, tmmn, def, aet, pdsi, soil, vs, srad	4638.3 m	monthly
NASA / USGS / JPL-Caltech	NASA SRTM Digital Elevation	USGS/SRTMGL1_003	elevation	30 m	-

Table 3 presents the features employed for predicting paddy yields, which have been associated with the image collection within the GEE. This mapping process is undertaken to streamline data acquisition, ensuring alignment between the collected data and the specified features.

Table 3. Mapping of features to image collection

Feature	Band	Unit	Image Collection
Elevation	elevation	m	USGS/SRTMGL1_003
Slope	slope	^{0}	USGS/SRTMGL1_003
Minimum temperature	tmmn	C	IDAHO_EPSCOR/TERRACLIMATE
Maximum temperature	tmmx	C	IDAHO_EPSCOR/TERRACLIMATE
Precipitation accumulation	pr	mm	IDAHO_EPSCOR/TERRACLIMATE
Wind-speed	vs	m/s	IDAHO_EPSCOR/TERRACLIMATE
Surface downward shortwave radiation	srad	W/m^2	IDAHO_EPSCOR/TERRACLIMATE
Soil moisture	soil	mm	IDAHO_EPSCOR/TERRACLIMATE
Actual evapotranspiration	aet	mm	IDAHO_EPSCOR/TERRACLIMATE
Palmer Drought Severity Index	pdsi	-	IDAHO_EPSCOR/TERRACLIMATE
Fraction of Photosynthetically Active Radiation	Fpar	-	MODIS/061/MCD15A3H
Leaf Area Index	Lai	-	MODIS/061/MCD15A3H
VV, VH Polarization	VV, VH	-	COPERNICUS/S1_GRD
Multispectral/Surface Reflectance	B1-12	nm	COPERNICUS/S2_SR_HARMONIZED
	SR_B1-7	µm	LANDSAT/LC08/C02/T1_L2
	sur_refl_b01, sur_refl_b02	nm	MODIS/061/MOD09Q1

The vegetation index is a calculated numerical value derived from specific spectral bands [12]. Table 4 provides an overview of various vegetation indices associated with crop yield predictions, along with the corresponding mathematical formulas employed.

Table 4. Vegetation Indices and their corresponding spectral bands and mathematical formulas

Name	Formula	Ref
NDVI	$\frac{NIR - Red}{NIR + Red}$	[39]
MSI	$\frac{SWIR1}{NIR}$	[40]
GNDVI	$\frac{NIR - Green}{NIR + Green}$	[41]
ExG	$2 \times Green - Red - Blue$	[42]
NDWI	$\frac{Green - NIR}{Green + NIR}$	[43]
VARI	$\frac{Green - Red}{Green + Red - Blue}$	[44]
PVR	$\frac{Green - Red}{Green + Red}$	[45]

5.2 Data Acquisition

The data acquisition process commences with the initial data collection phase, which relies on a combination of ground-truth data, factors influencing crop yields, and image retrieval from GEE. Ground-truth data is sourced from the Crop Cutting Survey, and the corresponding survey coordinates

serve as reference points for extracting image data from GEE. At this stage, the data generated consists of spectral band information obtained from the image collection, corresponding to specific coordinates and timestamps. Since the spectral data we gather originates from Sentinel-2 Level 2A and Landsat 8 Level 2, which have undergone radiometric and atmospheric corrections, additional corrections are unnecessary, thus conserving time and resources. Our primary task during this preprocessing phase is cloud and cloud shadow masking to ensure that the extracted band values accurately represent ground conditions. Subsequently, we proceed to calculate vegetation indices using established formulas. The outcome of this process encompasses vegetation indices and other relevant data that align with the specified features. Figure 3 illustrates the stepwise procedure of data acquisition.

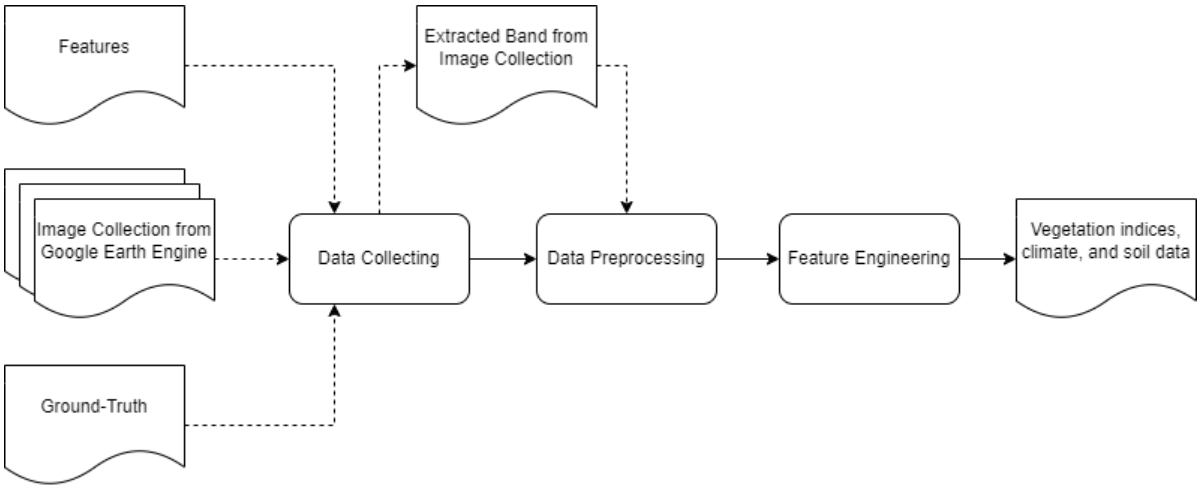

Figure 3. Data acquisition procedures

This research leveraged the capabilities of GEE, with a notable departure from traditional JavaScript-based console interaction. Instead, we harnessed the power of Python within the Google Colab environment to interface with GEE seamlessly. Subsequently, the data produced by GEE was efficiently stored on both Google Drive and local storage repositories. The architecture underpinning this data acquisition process is visually presented in Figure 4.

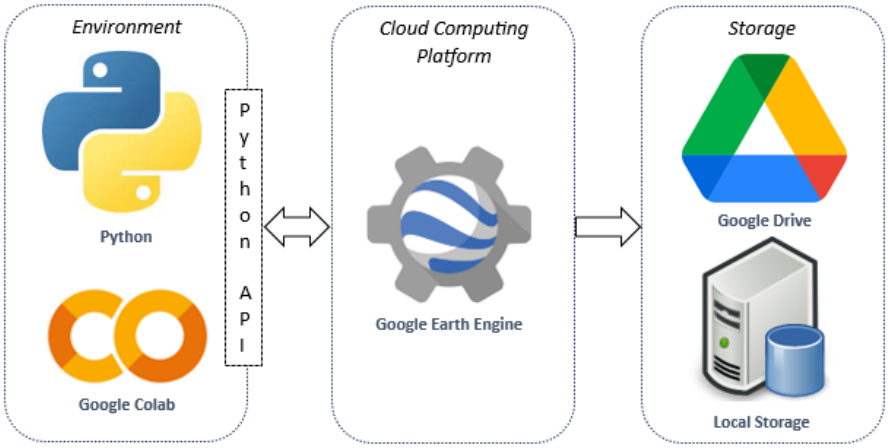

Figure 4. Data acquisition architecture

5.3 Dataset

5.3.1 Data Structure

The data obtained through the satellite image acquisition process is stored in a CSV and TIF file, generating a total of 42,948 data records from 306 coordinate points spanning the period of 2019-2020. Notably, 53.95% of the acquired data was marked as null, primarily due to the cloud and cloud shadow masking procedures. Detailed insights into the dataset's structure and metadata are presented in Figure 5.

```
latitude,longitude,date,ndvi,msi,gndvi,exg,msavi2,ndwi,evi,vari,pvr,mtvi,wdrvi,rgvi
-6.43537624,107.3890357,2019-11-06,0.6523,0.5013,0.5483,0.0311,0.3531,-0.5483,0.4725,0.3279,0.1619,0.3156,-0.0255,0.6307
-6.43537624,107.3890357,2019-11-11,0.6716,0.5205,0.5777,0.0276,0.3413,-0.5777,0.4367,0.2984,0.1534,0.2994,0.0089,0.6536
-6.43537624,107.3890357,2019-11-16,0.7940,0.4829,0.7095,0.0299,0.4582,-0.7095,0.5255,0.3393,0.1935,0.4091,0.2705,0.8036
-6.43537624,107.3890357,2019-11-21,0.4713,0.5717,0.3785,0.0210,0.2618,-0.3785,0.4418,0.2616,0.1130,0.2624,-0.2848,0.4232
-6.43537624,107.3890357,2019-12-01,0.8014,0.4784,0.7202,0.0320,0.4631,-0.7202,0.5169,0.3125,0.1919,0.4040,0.2893,0.8083
```

```
latitude,longitude,date,temp_min,temp_max,wind_speed,srad,precipitation,evapotranspiration,pdsi,soil_moisture
-6.43537624,107.3890357,2019-08-01,22.00,33.20,2.40,231.00,6.00,69.90,-2.80,142.90
-6.43537624,107.3890357,2019-09-01,22.70,34.00,2.70,253.70,5.00,50.50,-3.60,96.70
-6.43537624,107.3890357,2019-10-01,23.70,34.50,2.00,251.30,57.00,76.40,-3.90,74.30
-6.43537624,107.3890357,2019-11-01,23.80,33.40,1.30,248.00,105.00,108.10,-4.80,66.00
-6.43537624,107.3890357,2019-12-01,23.50,32.10,1.00,198.00,334.00,119.50,-4.30,263.60
```

```
latitude,longitude,id_loc,date,file
-6.43537624,107.3890357,117,2019-11-06,karawang/117_2019-11-06.tif
-6.43537624,107.3890357,117,2019-11-11,karawang/117_2019-11-11.tif
-6.43537624,107.3890357,117,2019-11-16,karawang/117_2019-11-16.tif
-6.43537624,107.3890357,117,2019-11-21,karawang/117_2019-11-21.tif
-6.43537624,107.3890357,117,2019-12-01,karawang/117_2019-12-01.tif
```

Figure 5. Structure of vegetation indices data (top), climate and soil data (middle), and RGB satellite images metadata (bottom)

Utilizing the derived vegetation index data, we conducted a straightforward analysis by plotting graphs depicting NDVI and EVI against date. The NDVI and EVI datasets were resampled using the median method with a monthly interval. The outcomes are illustrated in Figure 6. The graph shows a clear pattern, with NDVI displaying peak values indicative of the vegetative phases of paddy growth, while the valleys correspond to planting or harvesting periods. Notably, the data from Sentinel-2 consistently generates a clearer and more coherent pattern compared to Landsat 8. This discrepancy can be attributed to the superior spatial and temporal resolution capabilities of Sentinel-2 in contrast to Landsat 8.

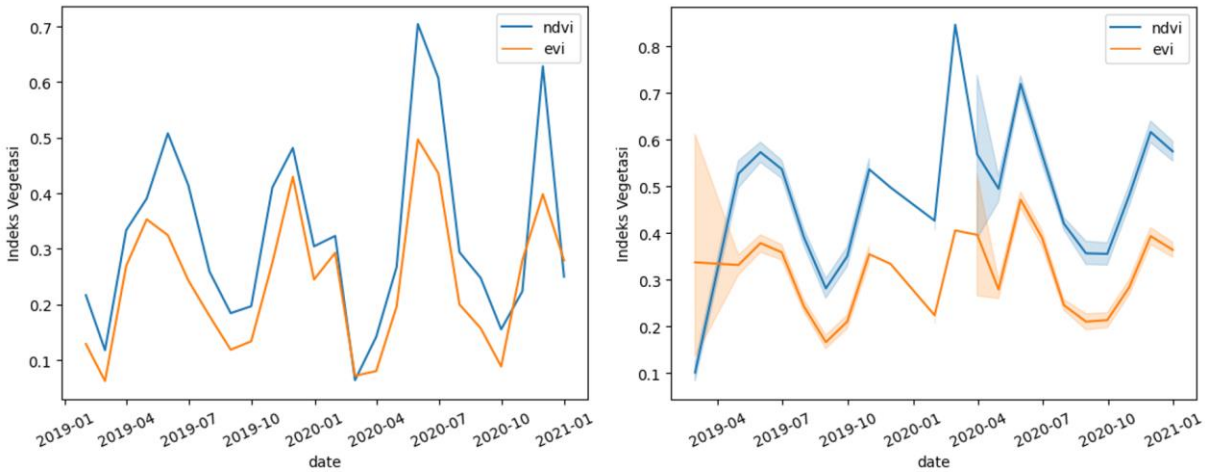

Figure 6. Comparison of NDVI and EVI between Sentinel-2 (left) and Landsat 8 (right)

5.4 Discussion

Numerous aspects can be explored concerning the establishment of a precision agriculture dataset utilizing satellite image data through GEE.

5.4.1 Area of Interest. The research area of interest in this study corresponds to the region covered by the Crop Cutting Survey. The projected paddy harvest productivity, as per official statistical data released by Statistics Indonesia [4], is quantified in quintals per hectare, prompting our adoption of a 1-hectare (10,000 square meters) boundary for our research area. For each existing coordinate point, a 1-hectare area is delineated. In Figure 7, we illustrate the obtained area, defined using a 100-meter buffer around Sentinel-2 and Landsat 8 satellite images. Sentinel-2 exhibits a 10-meter spatial resolution for the Red, Green, and Blue spectral channels, while Landsat 8 offers a 30-meter resolution. This disparity in resolution influences the resulting pixel area; Sentinel-2 satellite imagery represents a 10x10 pixel configuration, whereas Landsat 8 employs a 3x3 pixel arrangement. Consequently, there is a difference in the resulting area size, with Sentinel-2 encompassing 1 hectare, while Landsat 8 covers only 0.81 hectares.

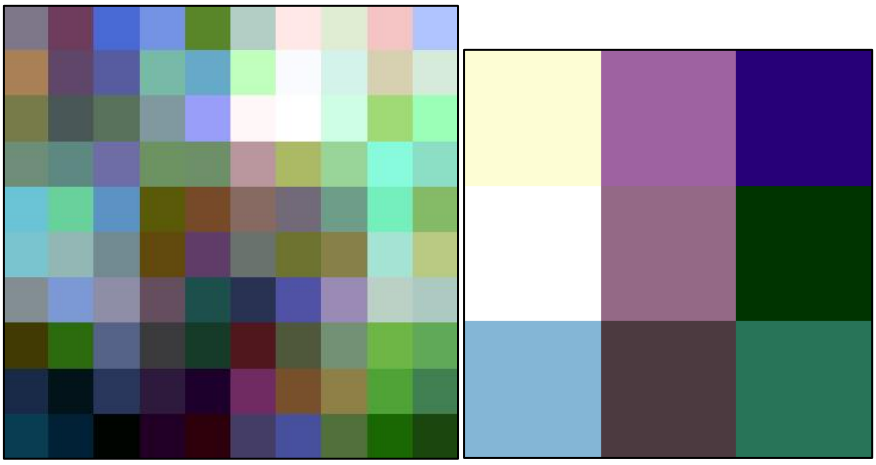

Figure 7. Comparison of pixel area between Sentinel-2 (left) and Landsat 8 (right)

5.4.2 Sequence of Data Processing

The sequence of data processing steps significantly impacts the computed vegetation index values. This pertains to the order of calculating the vegetation index formula and the subsequent resampling process. Table 5 presents the spectral values for a specific coordinate point in May 2019, while Table 6 showcases the NDVI and EVI values for May 2019 based on varying processing sequences. Upon examination, it becomes evident that variations in processing order lead to disparities in the calculated values.

Table 5. Spectral band values from Sentinel-2 for a selected coordinate point *

Date	Aerosol	Blue	Green	Red	RE1	RE2	RE3	NIR	RE4	Water Vapor	SWIR1	SWIR2
05-05-19	0.087	0.086	0.118	0.076	0.152	0.352	0.443	0.433	0.488	0.468	0.199	0.097
10-05-19	0.227	0.235	0.221	0.185	0.228	0.297	0.341	0.327	0.344	0.652	0.205	0.180
15-05-19	0.032	0.044	0.085	0.068	0.146	0.287	0.342	0.336	0.385	0.368	0.208	0.114
20-05-19	0.028	0.048	0.086	0.085	0.158	0.262	0.308	0.306	0.350	0.340	0.237	0.150
25-05-19	0.053	0.069	0.103	0.127	0.184	0.229	0.256	0.254	0.295	0.290	0.288	0.195
30-05-19	0.033	0.060	0.092	0.130	0.168	0.194	0.219	0.221	0.252	0.241	0.285	0.208

** longitude 107.5151737 and latitude -6.175384333 with 100-meter buffer*

Table 6. Variation in NDVI and EVI for May 2019 due to data processing order

Vegetation Indices	First Order		Score Difference
	Calculate VIs	Resampling	
NDVI	0.450	0.499	0.049
EVI	0.428	0.359	0.069

Despite the disparities in values, they do not impact the graphical representation. Figure 8 illustrates the graphs of NDVI and EVI when calculating the vegetation index first and resampling first.

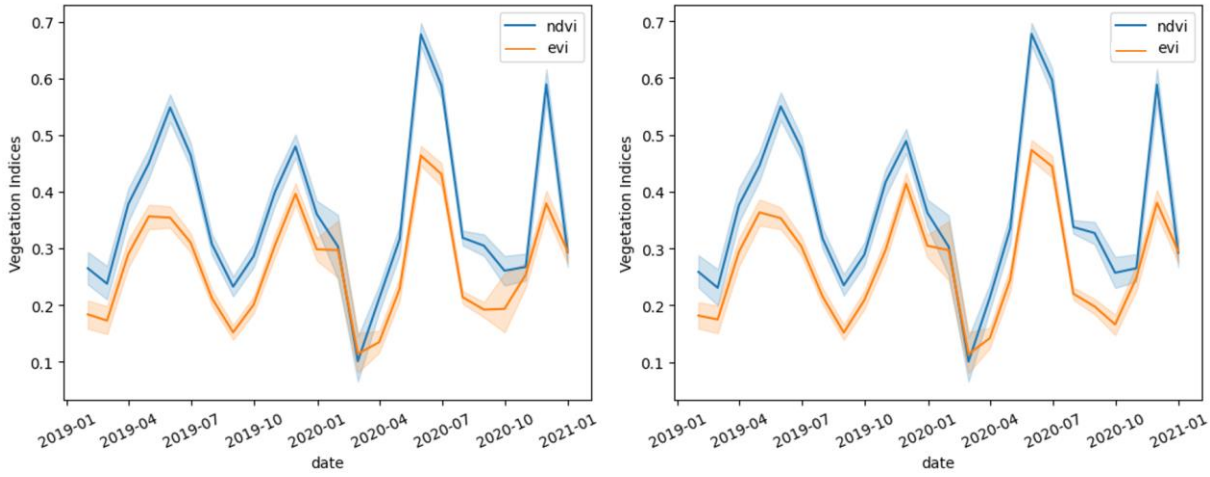

Figure 8. Comparison of NDVI and EVI between calculating vegetation indices first (Left) and resampling first (Right)

6. Conclusion

In this study, we analysed features relevant to the paddy crop and compared them across various image collections available in the Google Earth Engine (GEE) platform. Our data acquisition process was implemented using Python and Google Colab, guided by ground-truth data. During the data understanding phase, it became evident that the satellite imagery produced by Sentinel-2 outperformed that of Landsat 8. Notable considerations in crafting these precision agriculture datasets include defining the specific region of interest and establishing a structured sequence for data processing. Additionally, the presence of clouds and cloud shadows remains a common challenge associated with satellite imagery. While efforts were made to mask clouds and their shadows for enhanced data accuracy, the issue of data loss due to excessive cloud cover occasionally emerged. To address this, several techniques can be used to impute missing data [46], thereby complementing the study, and elevating the overall quality of the dataset.

References

[1] Sishodia RP, Ray RL, Singh SK, 2020, *Applications of Remote Sensing in Precision Agriculture: A Review*. Remote Sensing.

[2] FAO, 2021, *The State of Food Security and Nutrition in the World* Food and Agriculture Organization of the United Nations, Rome.

[3] Lobell D and Burke M, 2010, *On the Use of Statistical Models to Predict Crop Yield Responses to Climate Change* Agricultural and Forest Meteorology, vol. 150, no. 11, pp. 1443-1452.

[4] Badan Pusat Statistik, 2022, *Executive Summary of Paddy Harvested Area and Production in Indonesia 2021*, Jakarta.

[5] Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R., 2017, *Google Earth Engine: Planetary-scale geospatial analysis for everyone* Remote Sensing of Environment.

[6] Mulla D, 2013, *Twenty-five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps* Biosystems Engineering, vol. 114, no. 4, pp. 358-371.
[7] Shearer C, 2000, *The CRISP-DM Model: The New Blueprint for Data Mining*. Journal of Data Warehousing, 5, 13-22.
[8] Gumz, J., Fettermann, D.C., Frazzon, E.M., Kück, M. *Using Industry 4.0's Big Data and IoT to Perform Feature-Based and Past Data-Based Energy Consumption Predictions*. Sustainability 2022, 14, 13642. https://doi.org/10.3390/ su142013642
[9] Fuentes-Peñailillo, F., Carrasco Silva, G.; Pérez Guzmán, R., Burgos, I.; Ewertz, F. *Automating Seedling Counts in Horticulture Using Computer Vision and AI*. Horticulturae 2023, 9, 1134. https:// doi.org/10.3390/horticulturae9101134
[10] Huang Y, Chen Z, Xin Yu T, Huang Xzhi, Gu X, 2018, *Agricultural remote sensing big data: Management and applications*. In Journal of Integrative Agriculture (Vol. 17, Issue 9, pp. 1915–1931). Chinese Academy of Agricultural Sciences.
[11] Rembold F, Atzberger C, Savin I, Rojas O, 2013, *Using Low Resolution Satellite Imagery For Yield Prediction And Yield Anomaly Detection*. In Remote Sensing (Vol. 5, Issue 4, pp. 1704–1733).
[12] Xue J, Su B., 2017, *Significant Remote Sensing Vegetation Indices: A Review Of Developments And Applications*. In the Journal of Sensors
[13] Jaisyurahman U, Wirnas D, Trikoesoemaningtyas, Purnamawati H. 2019. *Dampak Suhu Tinggi terhadap Pertumbuhan dan Hasil Tanaman Padi*. J. Agron. Indonesia, 47(3):248-254.
[14] Khamid MBR, Junaedi A, Lubis I, Yamamoto Y, 2019, *Respon Pertumbuhan dan Hasil Padi (Oryza sativa L.) terhadap Cekaman Suhu Tinggi*. J. Agron. Indonesia, 47(2):119-125.
[15] Dulbari, Santoso E, Koesmaryono Y, Sulistyono E, 2018, *Pendugaan Kehilangan Hasil pada Tanaman Padi Rebah Akibat Terpaan Angin Kencang dan Curah Hujan Tinggi*. J. Agron. Indonesia, 46(1):17-23.
[16] Purwono, Dulbari, Santosa E, 2021, *Dampak Cuaca Ekstrim Terhadap Kehampaan Genotipe Padi: Pengantar Manajemen Produksi Berbasis Iklim*. J. Agron. Indonesia, 49(2): 136-146
[17] Cai Y, Guan K, Lobell D, Potgieter A B, Wang S, Peng J, Xu T, Asseng S, Zhang Y, You L, & Peng B. 2019. *Integrating Satellite and Climate Data to Predict Wheat Yield in Australia Using Machine Learning Approaches*. Agricultural and Forest Meteorology, 274, 144–159
[18] Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V. V., & Ciampitti, I. A. (2020). *Satellite-based Soybean Yield Forecast: Integrating Machine Learning and Weather Data for Improving Crop Yield Prediction in Southern Brazil*. Agricultural and Forest Meteorology.
[19] Islam, M. M., Matsushita, S., Noguchi, R., Ahamed, T. 2021. *Development of Remote Sensing-based Yield Prediction Models at The Maturity Stage of Boro Rice Using Parametric and Nonparametric Approaches*. Remote Sensing Applications: Society and Environment, 22.
[20] Fernandez-Beltran, R., Baidar, T., Kang, J., & Pla, F. 2021. *Rice-yield Prediction with Multi-Temporal Sentinel-2 Data and 3D CNN: A Case Study in Nepal*. Remote Sensing, 13(7).
[21] Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Zhang, J., Han, J., & Xie, J. 2021. *Integrating Multi-Source Data for Rice Yield Prediction across China using Machine Learning and Deep Learning Approaches*. Agricultural and Forest Meteorology, 297.
[22] Jeong, S., Ko, J., & Yeom, J. M. 2022. *Predicting Rice Yield At Pixel Scale Through Synthetic Use of Crop and Deep Learning Models with Satellite Data in South and North Korea*. Science of the Total Environment, 802.
[23] Ma, J. W., Nguyen, C. H., Lee, K., & Heo, J. 2018. *Regional-scale Rice-yield Estimation using stacked Auto-encoder with Climatic and MODIS data: a case study of South Korea*. International Journal of Remote Sensing.
[24] Yu, W., Yang, G., Li, D., Zheng, H., Yao, X., Zhu, Y., Cao, W., Qiu, L., & Cheng, T. (2023). *Improved prediction of rice yield at field and county levels by synergistic use of SAR, optical and meteorological data. Agricultural and Forest Meteorology*, 342, 109729. https://doi.org/10.1016/j.agrformet.2023.109729

[25] Ni, R., Tian, J., Li, X., Yin, D., Li, J., Gong, H., Zhang, J., Zhu, L., & Wu, D. (2021). *An enhanced pixel-based phenological feature for accurate paddy rice mapping with Sentinel-2 imagery in Google Earth Engine*. ISPRS Journal of Photogrammetry and Remote Sensing, 178, 282–296. https://doi.org/10.1016/j.isprsjprs.2021.06.018

[26] Fatchurrachman, Rudiyanto, Soh, N. C., Shah, R. M., Giap, S. G. E., Setiawan, B. I., & Minasny, B. (2023). *Automated near-real-time mapping and monitoring of rice growth extent and stages in Selangor Malaysia*. Remote Sensing Applications: Society and Environment, 31. https://doi.org/10.1016/j.rsase.2023.100993

[27] Rudiyanto, Minasny, B., Shah, R. M., Soh, N. C., Arif, C., & Setiawan, B. I. (2019). *Automated near-real-time mapping and monitoring of rice extent, cropping patterns, and growth stages in Southeast Asia using Sentinel-1 time series on a Google Earth Engine platform*. Remote Sensing, 11(14). https://doi.org/10.3390/rs11141666

[28] Rauf, U., Qureshi, W. S., Jabbar, H., Zeb, A., Mirza, A., Alanazi, E., Khan, U. S., & Rashid, N. (2022). *A new method for pixel classification for rice variety identification using spectral and time series data from Sentinel-2 satellite imagery*. Computers and Electronics in Agriculture, 193. https://doi.org/10.1016/j.compag.2022.106731

[29] Dale, D. S., Liang, L., Zhong, L., Reba, M. L., & Runkle, B. R. K. (2023). *Deep learning solutions for mapping contour levee rice production systems from very high resolution imagery*. Computers and Electronics in Agriculture, 211. https://doi.org/10.1016/j.compag.2023.107954

[30] Thorp, K. R., & Drajat, D. 2021. *Deep machine learning with Sentinel satellite data to map paddy rice production stages across West Java, Indonesia*. Remote Sensing of Environment, *265*

[31] Bem, P. P. de, de Carvalho Júnior, O. A., Carvalho, O. L. F. de, Gomes, R. A. T., Guimarães, R. F., & Pimentel, C. M. M. M. (2021). *Irrigated rice crop identification in Southern Brazil using convolutional neural networks and Sentinel-1 time series*. Remote Sensing Applications: Society and Environment, 24. https://doi.org/10.1016/j.rsase.2021.100627

[32] Pang, J., Zhang, R., Yu, B., Liao, M., Lv, J., Xie, L., Li, S., & Zhan, J. (2021). *Pixel-level rice planting information monitoring in Fujin City based on time-series SAR imagery*. International Journal of Applied Earth Observation and Geoinformation, 104. https://doi.org/10.1016/j.jag.2021.102551

[33] Xiao, D., Niu, H., Guo, F., Zhao, S., & Fan, L. (2022). *Monitoring irrigation dynamics in paddy fields using spatiotemporal fusion of Sentinel-2 and MODIS*. Agricultural Water Management, 263. https://doi.org/10.1016/j.agwat.2021.107409

[34] Cai, Y., Lin, H., & Zhang, M. (2019). *Mapping paddy rice by the object-based random forest method using time series Sentinel-1/Sentinel-2 data*. Advances in Space Research, 64(11), 2233–2244. https://doi.org/10.1016/j.asr.2019.08.042

[35] Singha, C., & Swain, K. C. (2023). *Rice crop growth monitoring with sentinel 1 SAR data using machine learning models in google earth engine cloud*. Remote Sensing Applications: Society and Environment, 32. https://doi.org/10.1016/j.rsase.2023.101029

[36] Suryono, H.; Kuswanto, H.; Iriawan, N. *Two-Phase Stratified Random Forest for Paddy Growth Phase Classification: A Case of Two-Phase Stratified Random Forest for Paddy Growth Phase Classification: A Case of Imbalanced Data*. https://doi.org/10.3390/10.3390/su142215252

[37] Xia, L., Zhao, F., Chen, J., Yu, L., Lu, M., Yu, Q., Liang, S., Fan, L., Sun, X., Wu, S., Wu, W., & Yang, P. (2022). *A full resolution deep learning network for paddy rice mapping using Landsat data*. ISPRS Journal of Photogrammetry and Remote Sensing, 194, 91–107. https://doi.org/10.1016/j.isprsjprs.2022.10.005

[38] Yang, R., Ahmed, Z. U., Schulthess, U. C., Kamal, M., & Rai, R. (2020). *Detecting functional field units from satellite images in smallholder farming systems using a deep learning based computer vision approach: A case study from Bangladesh*. Remote Sensing Applications: Society and Environment, 20. https://doi.org/10.1016/j.rsase.2020.100413

[39] Rouse, J.W., Haas, R.H., Schell, J.A. and Deering, D.W. 1973. *Monitoring Vegetation Systems in the Great Plains with ERTS (Earth Resources Technology Satellite).* Proceedings of 3rd Earth Resources Technology Satellite Symposium, Greenbelt, 10-14 December, SP-351, 309-317.
[40] Hunt E R, Rock B N, 1989, *Detection of changes in leaf water content using Near and Middle-Infrared reflectances*, Remote Sensing of Environment, Volume 30, Issue 1, Pages 43-54
[41] Gitelson AA, Kaufman YJ, Merzlyak MN, 1996, *Use of a green channel in remote sensing of global vegetation from EOS-MODIS*, Remote Sensing of Environment, Volume 58, Issue 3, Pages 289-298
[42] Woebbecke DM, Meyer GE, Von Bargen K, Mortensen DA. 1995. *Color indices for weed identification under various soil, residue, and lighting conditions*. Trans. ASAE 38, 259–269
[43] Zarco-Tejada PJ, Rueda CA, Ustin SL, 2003, *Water content estimation in vegetation with MODIS reflectance data and model inversion methods*, Remote Sensing of Environment, Volume 85, Issue 1, Pages 109-124,
[44] Gitelson AA, Merzlyak MN, Zur Y, Stark R, and Gritz U. 2001."*Non-Destructive and Remote Sensing Techniques for Estimation of Vegetation Status*". Papers in Natural Resources. 273
[45] Metternicht, G. 2003. *Vegetation indices derived from high-resolution airborne videography for precision crop management*. International Journal of Remote Sensing.
[46] Li J, Heap AD, 2013, *Spatial interpolation methods applied in the environmental sciences: A review*, Environmental Modelling & Software.