Optimizing Sampling Design for Chlorophyll-a Estimation in Inland Waters Using Sentinel-2 Imagery and Spatial Clustering

Azucena Pérez-Vega [1], Jean-François Mas [2], Jesús Delegido [3], Antonio Ruiz-Verdú [3]

[1] Departamento de Geomática e Hidráulica, Universidad de Guanajuato, Guanajuato, Mexico - azupv@ugto.mx
[2] Centro de Investigaciones en Geografía Ambiental, Universidad Nacional Autónoma de México, Morelia, Mexico - jfmas@ciga.unam.mx
[3] Cavanilles Institute of Biodiversity and Evolutionary Biology (ICBiBE), Universitat de València, Paterna, València, Spain

Keywords: Chlorophyll-a, Sentinel-2, empirical modeling, spatial clustering, water quality monitoring.

Abstract

Monitoring water quality in inland water bodies is critical for environmental management, yet traditional sampling methods are costly and spatially limited. Remote sensing offers a viable alternative by enabling large-scale assessment of water quality parameters such as chlorophyll-a (Chl-a) concentrations through empirical models linking spectral indices to *in situ* measurements. However, the accuracy of these models depends on representative sampling strategies that capture spatial and temporal variability. This study evaluates a clustering-based approach to optimize sampling site selection in the Solís Dam, Mexico, using Sentinel-2 imagery. We analyzed a one-year time series of Sentinel-2 data to compute spectral indices related to Chl-a and turbidity. Unsupervised K-means clustering was applied to stratify the reservoir into zones of distinct water quality variability, guiding the placement of 20 sampling sites. Field campaigns during dry and wet seasons (2024) provided Chl-a measurements, which were correlated with spectral indices. The Gi033BDA index showed the strongest correlation ($R^2 > 0.7$, $p < 0.01$) and was used to develop a linear regression model for Chl-a estimation. Results confirmed that clustering-derived sampling points effectively represented spatial variability, though temporal mismatches (1-day lag) and samples location inaccuracies introduced minor errors. The method demonstrates how pre-stratification using remote sensing can enhance sampling efficiency while maintaining model accuracy. This approach is particularly valuable for large-scale monitoring, reducing reliance on exhaustive field campaigns. Future work should address temporal dynamics and sensor resolution trade-offs for broader applicability.

1. Introduction

The assessment of water quality parameters such as chlorophyll concentration and turbidity in inland water bodies can be carried out using empirical models, combining field samples with multispectral imagery. Empirical models are based on statistical relationships between *in situ* data (e.g., chlorophyll or turbidity measurements) and spectral bands or indices derived from satellite or drone imagery (such as Sentinel-2, Landsat, or multispectral cameras). For example, linear or nonlinear regressions between reflectance in the visible and infrared bands and spectral indices such as the Normalized Difference Chlorophyll Index (NDCI) are used to estimate chlorophyll. Spectral water quality indices use specific combinations of spectral bands to assess water quality parameters such as turbidity, chlorophyll concentration, or dissolved organic matter (Beck et al., 2019; Chawla et al., 2020; Kallio, 2000). They are often focused on the detection of harmful algal blooms that can affect human health (Anderson et al., 2000; Linkov et al., 2009). It is also important to mention that free remote sensing data with high spatial and temporal resolution are currently available. There- fore, remote sensing is considered a viable and low-cost alternative for water quality monitoring (Chawla et al., 2020; Ritchie et al., 2003).

These methods allow efficient water quality monitoring, especially over large areas, reducing reliance on costly traditional sampling. However, they require validation with *in situ* data to ensure acceptable accuracy. To achieve this, an adequate sample size is essential. For example, Håkanson et al. (2007) highlights that typical monitoring or experimental programs may not collect sufficient data to reliably capture high variability, leading to unreliable results. Furthermore, a principle of regression analysis is that models should not be used to extrapolate beyond the data range from which they were constructed. It is therefore important to optimize sampling to ensure that sampling sites are able to represent all the variability of the parameter studied in the study area.

Atkinson et al. (2010) observed that random sampling is inefficient when data are spatially dependent. Spatial dependence means that observations that are close in space are more similar than those that are farther apart: they are related by a statistical correlation that generally decreases with increasing separation distance. Geostatistics can be used to analyze spatial data, explicitly accounting for spatial dependence. In particular, spatial correlation analysis, clustering, and kriging can be used to determine the required number of ground-based observations and the distance between them. For example, Di et al. (1989) used semivariograms to determine the number of samples and the appropriate sampling intervals to achieve desired levels of accuracy. Hawbaker et al. (2009) estimated forest vegetation structure and biomass using LiDAR data as a predictor and compared a random field sampling design with a stratified sampling design using LiDAR data as prior information. In the stratified sampling approach, they used aggregated LiDAR height measurements as prior information to stratify potential field sampling locations in our study area based on the mean and standard deviation of the LiDAR height. They found that the prediction errors of models built with a random sample were up to 68% higher than those of models built with a stratified sample.

Samples in remote sensing image classification are primarily

used as training data for classification models. Millard et al. (2015) analyzed the effects of input data characteristics on the Random Forest classification algorithm. Their results showed that the algorithm was highly sensitive to the training dataset. Lv et al. (2021) proposed a clustering-based sample selection method that applies histogram analysis to select more distinctive samples.

The objective of this study is to evaluate a sampling site selection method for fitting an empirical model relating field-measured chlorophyll concentration to spectral indices derived from multispectral images.

2. Study area

The Solís Dam is the largest reservoir in the state of Guanajuato, in central Mexico (Figure 1) and belongs to the Lerma-Toluca sub-basin. This sub-basin is among the most polluted in Mexico, a result of intense industrial and agricultural activities, as well as the presence of more than 3.5 million inhabitants (Cotler-Avalos et al., 2006).

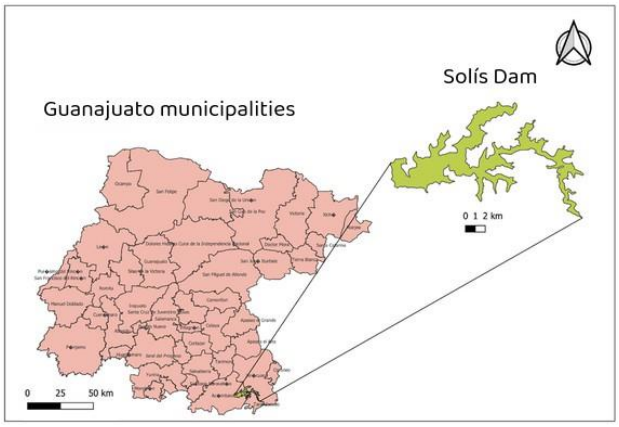

Figure 1. Location map of the Solís Dam in the state of Guanajuato. Source: Prepared by the authors.

3. Materials

Multispectral imagery from the Sentinel-2 satellite constellation was used. These images feature 12 bands from the visible and infrared with a spatial resolution of 10, 20, and 60 m depending on the spectral band (Table 1) and a temporal resolution of five days (less so from July 2024 when the calibration phase of Sentinel-2C, the third satellite in the Sentinel-2 constellation, ended). Pre-processed surface reflectance level (level 2A) images were obtained from the Copernicus Data Space Ecosystem (https://dataspace.copernicus.eu). Image processing and statistical analyses were performed using R software (R Core Team, 2025), specifically the CDSE packages (Karaman, 2025) for image acquisition, terra (Hijmans, 2025) for image preprocessing, waterquality (Johansen et al., 2023) for computing spectral water quality indices, and tmap (Tennekes, 2018) for mapping.

During field sampling, an HQ40D multiparameter probe and a 3-liter Van Dorn bottle were used. The samples were transported cold to the Soil and Water Analysis Laboratory at CIGA UNAM in Morelia. In the laboratory, a Metrohm 827 pH lab potentiometer, an OAKTON CON Series 510 conductivity meter,

Table 1. Spectral bands of the multi-spectral sensor (MSI) (Sentinel-2). The indices were calculated using the first nine bands.

Sentinel-2 Bands	Wavelength (μm)	Res (m)
Band 1 - Coastal aerosol	0.443	60
Band 2 - Blue	0.49	10
Band 3 - Green	0.56	10
Band 4 - Red	0.665	10
Band 5 - Red Edge	0.705	20
Band 6 - Red Edge	0.74	20
Band 7 - Red Edge	0.783	20
Band 8 - NIR	0.842	10
Band 8A - Red Edge	0.865	20
Band 9 - Water vapour	0.945	60
Band 10 - SWIR - Cirrus	1.375	60
Band 11 - SWIR	1.61	20
Band 12 - SWIR	2.19	20

a Metrohm 848 Tritino plus titrator, a 2100N turbidimeter, and a UV-Visible spectrometer were used.

4. Methods

In a first stage, satellite images were analyzed prior to fieldwork to determine the optimal location of the sampling points. Water samples were then collected and analyzed, and spectral indices were calculated based on the satellite images. The correlation between the values of different indices and chlorophyll concentration was analyzed, allowing the selection of the spectral indices that best represent this parameter. Finally, the representativeness of the sampling sites was evaluated, and sources of error such as spatial and temporal heterogeneity were identified (see Figure 2).

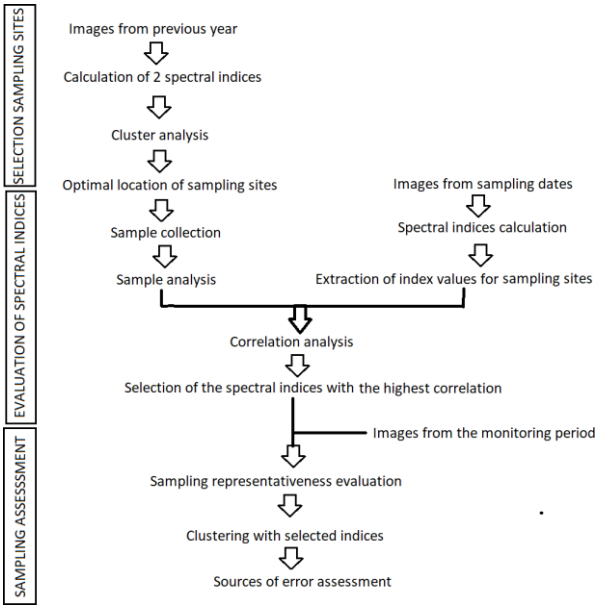

Figure 2. Flowchart of the main methodological steps

It is important that sampling best represents the spatiotemporal variability in water quality, i.e., some areas with high concentrations of chlorophyll and suspended solids, while others have low concentrations. However, in many cases, pre-sampling information is scarce, and the proposal here is to use remote sensing data to produce an *a priori* assessment of chlorophyll concentration distribution, allowing for stratification of sampling.

To this end, we sought to identify the areas with the greatest spectral variability, and possibly environmental variability, in relation to water quality parameters. To this end, all Sentinel-2 images with less than 40% cloud cover over the reservoir were selected for a one-year period prior to the field trips (May 1, 2023 to April 21, 2024), available in the Copernicus Data Space Ecosystem. The images were obtained using the R package CDSE and the monthly image with the least cloud cover over the water mirror was selected using the SCL (scene classification map) and QA layers.

For each monthly image, a spectral index related to chlorophyll concentration was calculated using the tri-band model of Dall'Olmo et al. (2003) modified by Sòria-Perpinyà et al. (2019), which can be calculated, in the case of Sentinel-2, using equations 1 and 2.

$$TBDO = b_6 \times \left(\frac{1}{b_4} - \frac{1}{b_5}\right), \quad (1)$$

$$[Chl - a] = 104.1 \times TBDO^2 + 221.1 \times TBDO + 2 \quad (2)$$

where TBDO = tri-band model
b_i = reflectance of band i of Sentinel-2
$[Chl - a]$ = estimate of the chlorophyll concentration in $\mu g/L$

Band 5 (red edge) was selected to represent turbidity, given its high degree of sensitivity to reflectance caused by material suspended in the water (Toming et al., 2016; Ogashawara et al., 2017). A stack of the 24 monthly images (12 related to turbidity and 12 to chlorophyll concentration) was created and unsupervised classification was applied using the K-means method (maximum number of iterations = 500, Lloyd's algorithm). The K-means method is a clustering algorithm that organizes data into K groups, assigning each point (here each pixel) to the group whose center (mean) is closest, iteratively adjusting the centers until they stabilize. It seeks to minimize the dispersion within each group, allowing the identification, in the present case, of areas with a similar spatiotemporal pattern in terms of chlorophyll concentration and turbidity. Twenty sampling sites were located to represent the different clusters.

4.1 Field Sampling and Analysis of Water Quality Parameters

Two field trips were conducted to sample the sites selected in the previous section during the dry and rainy seasons. The field samples were used to evaluate chlorophyll concentration (µg/L), turbidity (TNU), total suspended solids (mg/L), pH, electrical conductivity (µS), hardness (mg/L), alkalinity (mg/L), temperature, dissolved oxygen, and Secchi depth. Only the first parameter was used in this study.

4.2 Calculation and Selection of Spectral Indices

Based on Sentinel-2 images from the dates closest to sampling, the 22 water quality spectral indices available for Sentinel-2 were calculated in the R package waterquality. Additionally, the chlorophyll concentration index was calculated using the Dall-Olmo tri-band model (equation 1). One index of particular interest is Gi033BDA (Gitelson et al., 2003)(see equation 4), which is equivalent to TBDO (equation 1).

$$Gi033BDA = b_6 \times \left(\frac{1}{b_4} - \frac{1}{b_5}\right), \quad (3)$$

where Gi033BDA = spectral index
b_i = reflectance of band i of Sentinel-2

Pearson's correlation coefficients and their statistical significance were then calculated between each spectral index and the chlorophyll concentration obtained from the samples for the corresponding dates. A linear regression model was then fitted between chlorophyll concentration and the most correlated spectral indices, allowing chlorophyll concentration to be estimated from the satellite images. Once the spectral index(es) that best correlate with chlorophyll concentration were identified, we sought to evaluate whether the 20 sampling points were capable of representing the variability found in the images from dates close to the sampling dates.

Another way to evaluate the sampling point selection process is to perform a cluster analysis based on the spectral index(es) that were most correlated with the two water quality parameters and compare the cluster map with the one obtained with the two spectral indexes selected *a priori*.

Finally, we sought to evaluate two sources of error that could undermine the strength of the relationship between spectral indexes and chlorophyll concentration. The first is related to the uncertainty in the location of the sampling points and could lead to comparing a water sample taken at a certain point with the spectral information extracted from the image at different coordinates. This error could be due to errors in the GPS reading or to vessel movement during water sampling. To assess the possible effect of these spatial offsets, the estimated value of chlorophyll concentration in one image cell was compared with neighboring cells using special 3x3 and 5x5 kernel filters, evaluating offsets of approximately 20 and 40 m ($M_{3\times3}$ and $M_{5\times5}$).

$$M_{3\times3} = \begin{bmatrix} 1/8 & 1/8 & 1/8 \\ 1/8 & 0 & 1/8 \\ 1/8 & 1/8 & 1/8 \end{bmatrix}$$

$$M_{5\times5} = \begin{bmatrix} 1/16 & 1/16 & 1/16 & 1/16 & 1/16 \\ 1/16 & 0 & 0 & 0 & 1/16 \\ 1/16 & 0 & 0 & 0 & 1/16 \\ 1/16 & 0 & 0 & 0 & 1/16 \\ 1/16 & 1/16 & 1/16 & 1/16 & 1/16 \end{bmatrix}$$

The second source of error evaluated relates to temporal heterogeneity due to the dynamics of the water body. It is not always possible to obtain a cloud-free image on the same date as the water sampling, and changes in water quality parameters have been documented over short periods of time (Hunter et al., 2008; Deng et al., 2016; Xue et al., 2023). To assess the effect of these time lags, the difference between water concentration values estimated from pairs of images with collection dates that differ from several time periods was calculated.

5. Results

A clustering analysis (unsupervised classification) was carried out on ten clusters of the monthly images of the two spectral

indices related to chlorophyll concentration and turbidity, selected *a priori*. Twenty sites were selected to represent the diversity of values for the chlorophyll and sediment concentration indices represented by the different clusters (Fig. 3). To meet this requirement, these sites were distributed throughout the main body and arms of the reservoir, forming a type of transect from the reservoir inlet to the dam curtain. Figure 4 shows the area of each cluster and the number of points corresponding to each cluster. No sampling points were assigned to cluster 8, which corresponds to areas close to the shore which are out of water when the dam level is not at its maximum.

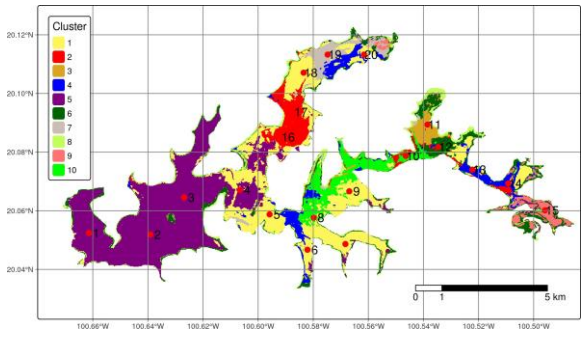

Figure 3. Distribution of the clusters and the sampling points.

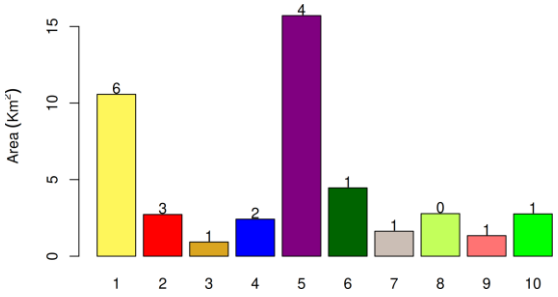

Figure 4. Area of each cluster of number of corresponding samples

Two sampling sessions were conducted on June 3 (dry season) and September 21, 2024 (end of the rainy season). Two Sentinel-2 images were obtained, taken on June 2 and September 20, 2024. The sampling and image collection dates do not coincide perfectly; there is a one-day time lag between them. The correlation between the spectral indices obtained from the Sentinel-2 images and the water quality parameters obtained from the water samples was evaluated.

The spectral index with the highest correlation with chlorophyll concentration is Gi033BDA (above 0.7 with a p-value < 0.01). The spectral indices used to create the cluster map show a lower correlation than Gi033BDA (Table 2). A linear regression was fitted between chlorophyll concentration and Gi033BDA, allowing chlorophyll concentration to be estimated from satellite images.

Next, we sought to evaluate whether the 20 sampling points selected based on the cluster map represented the variability

Parameter	Spectral indices		
	Gi033BDA	[*Chl-a*]	b_5
in situ [*Chla*]	0.75 ***	0.74***	0.33

Table 2. Correlation between the spectral indices obtained from the Sentinel-2 images and *in situ* chlorophyll concentration. The number of stars corresponds to the power of evidence against the null hypothesis, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, no star $p >= 0.05$

found in the Gi033BDA index value derived from the images that coincided with the sampling. As can be seen in Figures 5 and 6, the Gi033BDA index values for the sampling points correctly represent the variability found in the dam image set for both dates, with the exception of the most extreme outlier values.

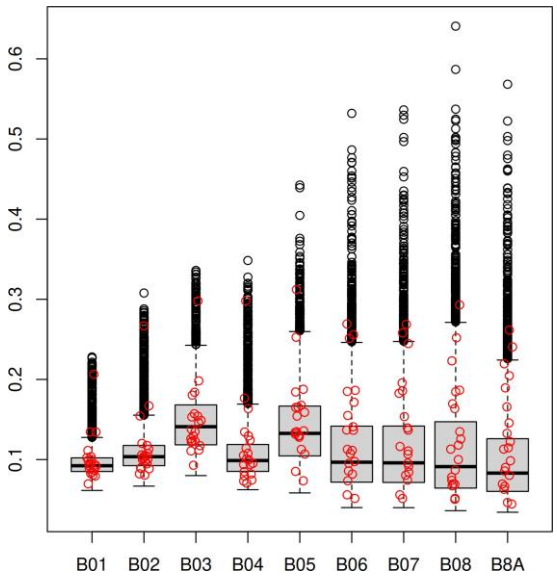

Figure 5. Gi033BDA index values in the Sentinel-2 image dated 02/06/2024 and at the sampling points (red circles)

To evaluate the sampling point selection process, a cluster analysis was performed using the Gi033BDA index. As can be seen in Figure 7, the overall distribution of the clusters is similar, although the overlap between the clusters in the two maps is only 60%. However, the distribution of the clusters along the dam and its branches would lead to a similar selection of sampling points.

Figure 8 represents the estimated chlorophyll concentration value calculated from the 02/06/2024 image. Significant spatial changes in values can be observed due to the presence of algal blooms.

Spatial filtering enables us to calculate, for each cell in the raster of Chlorophyll concentration, the difference between the value at a cell and the average value at a lag distance. The calculation allows us to evaluate the consequences of a shift between the sample location and the location used to extract the spectral value in the image. The 3x3 and 5x5 correspond approximately with lag distances of 20 and 40 m. Figures 9 and 10 illustrate the effect of uncertainty in the location of the sampling points, considering offsets of 20 and 40 m, respectively.

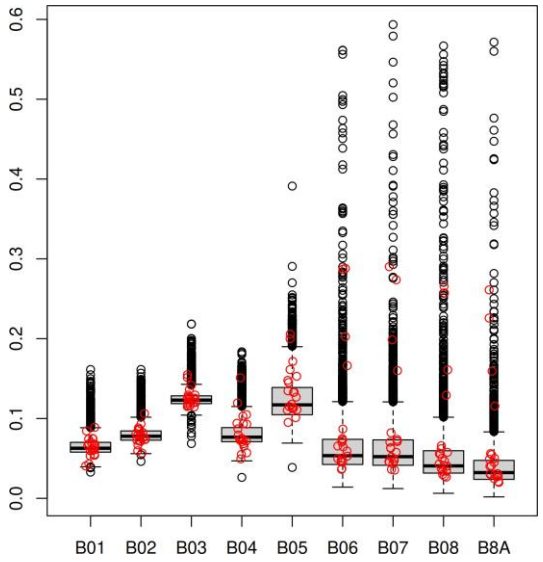

Figure 6. Gi033BDA index values in the Sentinel-2 image dated 20/09/2024 and at the sampling points (red circles)

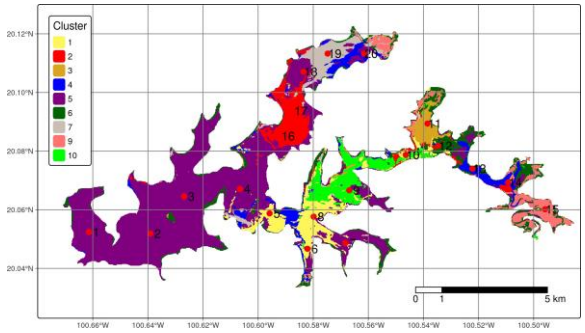

Figure 7. Distribution of the clusters based on Gi033BDA index and the sampling points.

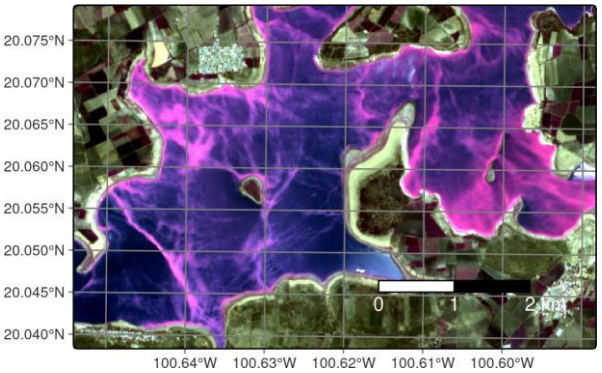

Figure 8. Chlorophyll concentration estimated from the Sentinel-2 image dated 02/06/2024.

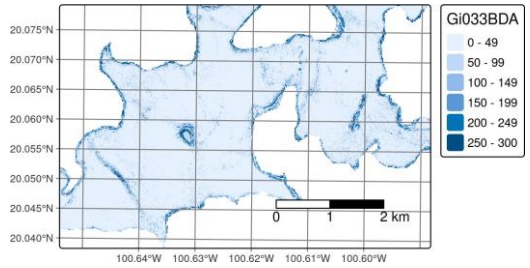

Figure 9. Assessment of the error in chlorophyll concentration associated with a 20-meter lag.

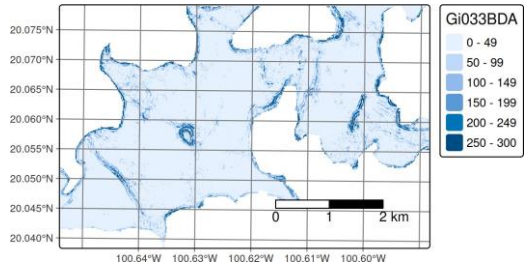

Figure 10. Assessment of the error in chlorophyll concentration associated with a 40-meter lag.

We computed the concentration of Chlorophyll from 163 Sentinel- 2 images taken from January 2023 to August 2025. Then we compared pairs of images, computing the absolute value of the cell-by-cell difference in chlorophyll concentration, and we cal- culated the mean value of the difference image. The boxplot (Figure 12) illustrated the mean value of the difference for the lag time between the dates of acquisition of the two compared images. For a lag of two days, the difference in chlorophyll concentration is generally around 20 µg/l . However, these val- ues are averaged for the entire water body and can be higher at specific locations.

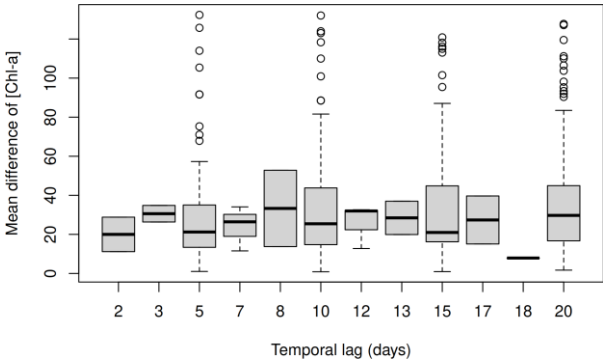

Figure 11. Effect of temporal lag in mean value of chlorophyll concentration.

For instance, Figure 12 shows the difference in chlorophyll con- centration between 19th and 21st March 2025, a two-day lag time. We can observe that during these 48 hours, certain areas present an increase of more than 20 µg/l and others a decrease reaching 60 µg/l, these drastic changes are related to the chan- ging patterns of algae blooms.

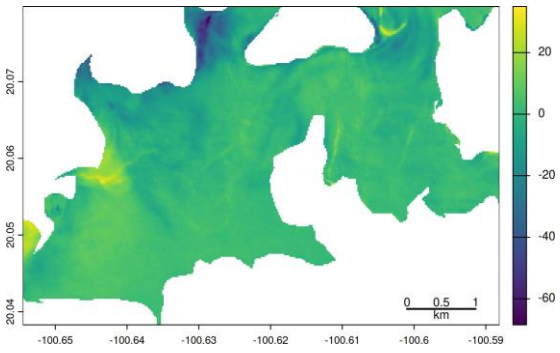

Figure 12. Difference in chlorophyll concentration between the 19th and 21st March 2025

6. Discussion

The significant fluctuations in chlorophyll concentration observed in the time series of satellite images and field data demonstrate the merit of stratifying the water body to optimize sampling. The 20 sampling sites selected based on unsupervised classi- fication (clustering) of a time series of images prior to sample collection adequately represented the dam's variability, allow- ing for a satisfactory fit of the linear regression model.

We did not find a similar approach in the literature for selecting sampling points in a water body; however, we saw several articles that describe different strategies for selecting sampling points at a regional level. For instance, Hedger et al. 2001 used Landsat imagery to compute a chlorophyll index, reflecting variations in chlorophyll a concentration, and compared two sampling strategies: a random sampling, which ignores spatial dependence between observations, and a systematic sampling, which exploits spatial dependence. They found that, for a given sample size, the systematic scheme provokes less error than the random sampling; and for a given error, the systematic scheme required smaller sample sizes than the random one. Catherine et al. 2008 proposed a method for selecting representative waterbodies to conduct a regional assessment of cyanobacteria distribution in the Île-de-France region (France). They applied a stratified sampling strategy to waterbodies based on ten groups of hydrographical zones, defined by their anthropogenic and geomorphological characteristics. Alilou et al. 2018; 2019 conducted a multi-criteria evaluation that incorporated the analytic network process, land use/change modeling, and a potential pollution assessment to prioritize sampling points and rationalize the water quality monitoring network in the Khoy watershed in northwest Iran.

The availability and characteristics of the Sentinel-2 data are particularly attractive for developing a water quality monitoring system for the Solís Dam: they are free, easily accessible online, and have a temporal resolution that allows for precise temporal monitoring (6 images per month, although during the rainy season the number of reusable images may be considerably lower). The spatial resolution, between 10 and 20 m for the bands used in this type of study, allows for the precise identification of localized phenomena such as algal blooms. Given the presence of highly localized chlorophyll concentration patterns (algal blooms), a high correlation between *in situ* measurements and spectral information depends on high spatial resolution and the simultaneity between field data collection and image acquisition. However, a model based on lower spatial and higher temporal resolution data, such as MODIS, can be of great interest, as the level of detail associated with high spatial resolution is not required for relatively large reservoirs. For example, Hu et al. (2010) and Shi et al. (2016) utilized MODIS to monitor cyanobacteria blooms in Lake Taihu, China, which has a surface area exceeding 2,000 km^2. Blix et al. 2018 used the Ocean and Land Color Instrument (OLCI) sensor onboard Sentinel-3 satellite (300 m spatial resolution) to estimate chlorophyll-a concentration, colored dissolved organic matter, and Total Suspended Matter, on Lake Balaton, whose surface area is 596 km^2. Long et al. 2025 utilized an unmanned aerial vehicle (UAV) to evaluate the water quality in Zhangshan Reservoir, a small inland reservoir (0.34 km^2) located in Chuzhou, Anhui, China.

The stratification strategy we proposed in the present study can be applied using any remotely sensed data, including multispectral satellite and UAV imagery. In the case of UAV images, data can be limited to a sufficient number of flight lines over the water body, eliminating the need for full coverage.

Other sources of error were identified that could lower the fit between *in situ* measurements and satellite image information. The first is the spatial lag between the location of *in situ* data collection and the extraction of spectral information. This source of error can be significant when there are highly localized high concentration patterns, such as bloom areas. The second is the concentration changes that can occur between *in situ* sampling and imaging. These changes can affect studies based on sampling and imaging dates that are different by one day, as in our case, but also by temporal differences of a few hours. Xue et al. (2023) show how wind and sunlight affect, in a very short time, the vertical movements of colonies of Microcystis, the dom- inant genus of cyanobacteria in eutrophic lakes. During field trips, it was observed how the wind that generally rises in the afternoon changed the distribution of the algae.

7. Conclusions

Field data collection requires a considerable amount of effort, time, and money. This collection is subject to inaccuracy due to human, instrument, or laboratory analysis errors. Using remote sensing and spatial statistical models, sampling effort can be significantly reduced. Mapping and assessing water quality parameters requires the integration of field, GIS, and remote sensing data using the most appropriate statistical methods. This work demonstrates that it is possible to reduce field data collection and to optimize data used in empirical models while maintaining accurate estimates of chlorophyll concentrations in inland water bodies.

In future research, we will address the spatiotemporal dynamics of water bodies and their analysis through sensors of different temporal and spatial resolutions.

Acknowledgements

This study was carried out in the scope of the projects 1) *Análisis de la Calidad, Cantidad, Gestión y Políticas Públicas ante escenarios de Cambio Climático en la Presa Solís, Guanajuato* with the support of the Dirección de Apoyo a la Investigación y el Posgrado, Universidad de Guanajuato and, 2) PAPIIT-UNAM project (IN112823) entitled: *Azolvamiento y eutroficación en*

presas periurbanas de zonas templadas de México: contribuciones para su evaluación y prospección (DGAPA-UNAM). We appreciate the processing of the water samples taken at the Soil and Water Laboratory (LASA) del Centro de Investigaciones en Geografía Ambiental at the Universidad Nacional Autónoma de México, work coordinated by Rosaura Páez. We thank Cinthia Casandra Sánchez Quintero, Yesenia Elena Trejo Nila, Luisa Fernanda Chavez Zamora and Ivan Alexandro Cervantes Esquivel for their support in the field and laboratory work. This study was completed during a stay at the University of Valencia of the first author with the support of the Department of Geomatics and Hydraulics at the University of Guanajuato.

References

Alilou, H., Moghaddam Nia, A., Saravi, M. M., Salajegheh, A., Han, D., Bakhtiari Enayat, B., 2019. A novel approach for se- lecting sampling points locations to river water quality monitor- ing in data-scarce regions. *Journal of Hydrology*, 573, 109-122. ttps://doi.org/10.1016/j.jhydrol.2019.03.068.

Alilou, H., Nia, A. M., Keshtkar, H., Han, D., Bray, M., 2018. A cost-effective and efficient framework to determine water qual- ity monitoring network locations. *The Science of the total en- vironment*, 624, 283-293. 10.1016/j.scitotenv.2017.12.121.

Anderson, D., Kaoru, Y., White, A., 2000. Estimated annual economic impacts from harmful algal blooms (habs) in the united states. Technical report, Woods Hole Oceanographic In- stitution.

Atkinson, P. M., Foody, G. M., Curran, P. J., Boyd, D. S., 2010. Assessing the ground data requirements for regional scale re- mote sensing of tropical forest biophysical properties Assessing the ground data requirements for regional scale remote sensing of tropical forest biophysical properties. *International Journal of Remote Sensing*, 21, 2571-2587.

Beck, R., Xu, M., Zhan, S., Johansen, R., Liu, H., Tong, S., Yang, B., Shu, S., Wu, Q., Wang, S., Berling, K., Murray, A., Emery, E., Reif, M., Harwood, J., Young, J., Nietch, C., Macke, D., Martin, M., Stillings, G., Stumpf, R., Su, H., Ye, Z., Huang, Y., 2019. Comparison of satellite reflectance algorithms for es- timating turbidity and cyanobacterial concentrations in product- ive freshwaters using hyperspectral aircraft imagery and dense coincident surface observations. *Journal of Great Lakes Re- search*, 45, 413-433.

Blix, K., Pálffy, K., R. Tóth, V., Eltoft, T., 2018. Remote Sensing of Water Quality Parameters over Lake Balaton by Using Sentinel-3 OLCI. *Water*, 10(10). https://www.mdpi.com/2073- 4441/10/10/1428.

Catherine, A., Troussellier, M., Bernard, C., 2008. Design and application of a stratified sampling strategy to study the regional distribution of cyanobacteria (Ile- de-France, France). *Water Research*, 42(20), 4989-5001. https://doi.org/10.1016/j.watres.2008.09.028.

Chawla, I., Karthikeyan, L., Mishra, A. K., 2020. A review of remote sensing applications for water security: Quantity, qual- ity, and extremes. *Journal of Hydrology*, 585, 124826.

Cotler-Avalos, H., Mazari-Hiriart, M., Anda-Sánchez, J., 2006. Atlas de la cuenca lerma-chapala. Technical report, SEMARNAT, INE-UNAM, Instituto de Ecolog´ıa.

Dall'Olmo, G., Gitelson, A. A., Rundquist, D. C., 2003. Towards a unified approach for remote estimation of chlorophyll-a in both terrestrial vegetation and turbid productive waters. *Geophysical Research Letters*, 30.

Deng, J., Chen, F., Liu, X., Peng, J., Hu, W., 2016. Horizontal migration of algal patches associated with cyanobacterial blooms in an eutrophic shallow lake. *Ecological Engineering*, 87, 185-193.

Di, H., Kemp, R., Trangmar, B., 1989. Use of Geostatistics in Designing Sampling Strategies for Soil Survey. *Soil Science Society of America Journal*, 53, 1163-1167.

Gitelson, A. A., Gritz, Y., Merzlyak, M. N., 2003. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *Journal of Plant Physiology*, 160, 271-282.

Hawbaker, T. J., Keuler, N. S., Lesak, A. A., Gobakken, T., Contrucci, K., Radeloff, V. C., 2009. Improved estimates of forest vegetation structure and biomass with a LiDAR-optimized sampling design. *J. Geophys. Res*, 114, 0-04.

Hedger, R. D., Atkinson, P. M., Malthus, T. J., 2001. Optimizing sampling strategies for estimating mean water quality in lakes using geostatistical techniques with remote sensing. *Lakes & Reservoirs: Science, Policy and Management for Sustainable Use*, 6(4), 279-288. https://doi.org/10.1046/j.1440-1770.2001.00159.x.

Hijmans, R. J., 2025. terra: Spatial data analysis. R package version 1.8-60.

Hu, C., Lee, Z., Ma, R., Yu, K., Li, D., Shang, S., 2010. Moderate Resolution Imaging Spectroradiometer (MODIS) observations of cyanobacteria blooms in Taihu Lake, China. *China, J. Geophys. Res*, 115, 4002.

Hunter, P. D., Tyler, A. N., Willby, N. J., Gilvear, D. J., 2008. The spatial dynamics of vertical migration by Microcystis aeruginosa in a eutrophic shallow lake: A case study using high spatial resolution time-series airborne remote sensing. *Limnology and Oceanography*, 53, 2391-2406.

Håkanson, L., 2007. A Data Reduction Exercise to Detect Threshold Samples for Regression Models to Predict Key Water Variables. *International Review of Hydrobiology*, 92(1), 84-97.

Johansen, R., Nowosad, J., Reif, M., Emery, E., 2023. waterquality: Satellite derived water quality detection algorithms. R package version 1.0.0.

Kallio, K., 2000. *Remote Sensing as a Tool for Monitoring Lake Water Quality*. John Wiley & Sons, 237–240.

Karaman, Z., 2025. Cdse: Copernicus data space ecosystem api wrapper.

Linkov, I., Satterstrom, F. K., Loney, D., Steevens, J. A., 2009. The impact of harmful algal blooms on usace operations. Technical report, US Army Engineer Research and Development Center.

Long, C., Zhang, J., Xia, X., Liu, D., Chen, L., Yan, X., 2025. High-Resolution Water Quality Monitoring of Small Reservoirs Using UAV-Based Multispectral Imaging. *Water*, 17(11). https://doi.org/10.3390/w17111566.

Lv, Z., Li, G., Yan, J., Benediktsson, J. A., You, Z., 2021. Training Samples Enriching Approach for Classification Improvement of VHR Remote Sensing Image. *IEEE Geoscience and Remote Sensing Letters*, 19, 2022.

Millard, K., Richardson, M., 2015. On the Importance of Training Data Sample Selection in Random Forest Image Classification: A Case Study in Peatland Ecosystem Mapping. 7, 8489-8515.

Ogashawara, I., Mishra, D. R., Gitelson, A. A., 2017. Remote Sensing of Inland Waters: Background and Current State-of-the-Art. *Bio-optical Modeling and Remote Sensing of Inland Waters*, 1-24.

R Core Team, 2025. R: A language and environment for statistical computing.

Ritchie, J. C., Zimba, P. V., Everitt, J. H., 2003. Remote Sensing Techniques to Assess Water Quality. *Photogrammetric Engineering and Remote Sensing*, 69, 695-704.

Shi, K., Zhang, Y., Zhou, Y., Liu, X., Zhu, G., Qin, B., Gao, G., 2016. Long-term MODIS observations of cyanobacterial dynamics in Lake Taihu: Responses to nutrient enrichment and meteorological factors OPEN. *Nature Publishing Group*.

Sòria-Perpinyà, X., Urrego, P., Pereira-Sandoval, M., Ruiz-Verdú, A., Soria, J. M., Delegido, J., Moreno, J., 2019. Monitoring the ecological state of a hypertrophic lake (Albufera of València, Spain) using multitemporal Sentinel-2 images. *Limnetica*, 38, 457-469.

Tennekes, M., 2018. tmap: Thematic Maps in R. *Journal of Statistical Software*, 84, 1-39. https://doi.org/10.18637/jss.v084.i06.

Toming, K., Kutser, T., Laas, A., Sepp, M., Paavel, B., Nõges, T., 2016. First Experiences in Mapping Lake Water Quality Parameters with Sentinel-2 MSI Imagery.

Xue, K., Ma, R., Shen, M., Wu, J., Hu, M., Guo, Y., Cao, Z., Xiong, J., 2023. Horizontal and vertical migration of cyanobacterial blooms in two eutrophic lakes observed from the GOCI satellite. *Water Research*, 240, 120099.