Machine learning models for satellite-based coral reef mapping

Teo Nguyen

UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR,
ÉCOLE DOCTORALE 211
Laboratoire de Mathématiques et de leurs Applications de Pau (LMAP)
&
MACQUARIE UNIVERSITY
School of Mathematical and Physical Sciences

Machine learning models for satellite-based coral reef mapping

---

Teo Nguyen

A thesis submitted for the degree of
Doctor of Philosophy in Mathematics

Thesis succesfully defended the 13th December 2023 in front of the jury composed by:

Christian PAROISSIN	Associate Professor	UPPA	President
Audrey MINGHELLI	Professor	UTLN	Reviewer
Anne RUIZ-GAZEN	Professor	TSE	Reviewer
Benoit LIQUET	Professor	MQ/UPPA	Co-supervisor
Kerrie MENGERSEN	Distinguished Professor	QUT	Supervisor
Sarat MOKA	Doctor	MQ/UNSW	Co-supervisor
Damien SOUS	Associate Professor	UPPA/UTLN	Supervisor
Nan YE	Doctor	UQ	Examiner

Ce qui n'allait pas j'ai pris du temps pour le changer,
J'ai pris du temps, ça prend du temps d'essayer.
Ben Mazué

Acknowledgements

I would like to offer my profound respects to the First Nations people and their lands which were stolen from them two hundred years ago by us, the European, a transgression that continues to this day. I recognize that I have walked upon their sacred grounds, and I carry both regret and gratitude for this privilege. It is my hope that every individual reading these words will reflect upon the profound injustice we inflicted to these communities. We destroyed the lives of hundreds of thousands people who had thrived in harmony with nature for hundreds of thousands years. May their wisdom be one day understood by all of us. May the day come where we give them back their lands and let them live peacefully with nature. May we learn from them to cherish and protect nature, valuing its precious resources without perpetuating the cycle of overexploitation and destruction that has persisted for far too long.

First of all, I would like to thank all of my supervisors. I would like to express my gratitude towards **Benoit Liquet**, who trusted me and spontaneous proposed me this PhD at a time where I had stopped thinking about going into the academic field. Without him, I am absolutely certain I would never have done a PhD. I would like to acknowledge his patience, his knowledge about statistics (and sports, even though it is less necessary to my PhD), and his support during the online weekly meeting when we had 10 hours time difference as well as the meetings in-person, almost everyday, in Macquarie University. I would like to thank **Damien Sous** for his trust, his rigor, his writing skills, for the speed at which he could become available when I had any question, and for framing me in a way that makes me more efficient in my work. I would like to thank **Kerrie Mengersen** for her huge knowledge, for her availability during our weekly meetings and for her capability to solve in five minutes some problems I had had for days, or weeks. Finally, I would like to thank **Sarat Moka** for his help, his knowledge I was lacking in algebra, for the three-people meetings we had in Macquarie, and for his presence at my first conference in Darwin.

I would also like to thank **Samuel Meulé** for his time and help with Maupiti data; without him I would have spent half my thesis trying to figure out how QGIS works. More generally, I would like to thank the whole GLADYS team for the Maupiti dataset.

I would like to thank **Baptiste** and **Carla**, two brilliant interns, that helped me solve some problems of this thesis, and saved me a lot of time.

I am also grateful to the cheerful people who welcomed me on the 6th floor of 12 Wally's Walk in Macquarie University, and with whom I shared some meals (and/or beers).

Then, I would like to extend my gratitude to the PhD students or postdoc with whom I've had deep interesting (or even shallow) conversations, in UPPA or in Macquarie: **Aurélien**,

Bastien, **Claire**, **Floren**, **Ibrahim**, **Sébastien**. A special thanks to Bastien and Claire with whom I've shared my office; I know I have been quite an absent roomate but I truly enjoyed our time together and all your advices (about academia, or life in general).

Finally, I would like to thank all my relatives.

Tout d'abord, merci à mes amis, que je ne nommerai pas tous, d'avoir toujours été à mes côtés. Une première mention spéciale (car je suis obligé) à **Guillaume** qui a, une fois, tenté de m'aider avec une équation, mais dont la tentative n'a évidemment pas abouti puisque son niveau en maths n'a d'égal que son pauvre niveau d'anglais ;). Deuxième mention spéciale pour **Antho**, **Rémi** et **Romain**, avec lesquels j'ai grandi et je grandis encore.

Évidemment, je remercie **Camille**. Merci pour beaucoup de choses. Merci d'avoir été tout le temps là, même à l'autre bout du monde, pour vivre ensemble ces moments magnifiques. Grâce à toi j'ai vécu et je vis ma meilleure vie.

Enfin, j'ai la chance d'être extrêmement bien entouré par une famille géniale, et d'avoir baigné dans un environnement propice à mon épanouissement. Je remercie tous mes cousins et cousines, qui ont été d'une superbe compagnie sur la côte landaise. Je remercie ma soeur, ma deuxième mère, pour sa présence et sa patience, pour m'avoir fait grandir. Je remercie mon frère, ce "leader charismatique", pour m'avoir toujours soutenu et aidé dans mes études, pour les rires, pour toutes les qualités que je n'ai pas. Je remercie maman et papa pour leur amour, leurs sacrifices, pour avoir toujours fait de leur mieux pour nous donner le meilleur possible. Pour terminer, j'ai une tendre pensée pour mamie Dédée, qui n'aura pas pu assiter à la fin de cette thèse mais dont je ne doute pas qu'elle aurait été très fière; on ne t'oublie pas.

Abstract

The ongoing crisis of climate change necessitates the development of effective methods for monitoring and mapping environmental features and species to ensure their preservation. This thesis explores the application of machine learning algorithms to efficiently map coral reefs using multispectral satellite images. The Maupiti lagoon in French Polynesia serves as a case study. The research led to the production of an automated tool capable of generating coral reef maps from satellite images. Moreover, the tool can be adapted to map other ecosystems, such as forests or ice sheets, provided that the model is retrained with relevant data.

To begin, a comprehensive literature review investigates current methods and trends in utilizing machine learning algorithms for coral reef mapping. Then, the attempts to develop the tool led us to face the special case of compositional data, which are data carrying relative information and lying in a mathematical space known as simplex. Adaptations of conventional methods are required to address the specific characteristics of this space.

First, in response to data imbalance, an oversampling technique is developed specifically for compositional data. Additionally, a spatial autoregressive model based on the Dirichlet distribution is formulated to account for spatial effects that may arise in the mapping process.

Finally, we present the implementation of our final mapping tool. To achieve the desired objective, a two-staged classification process is implemented, combining pixel-based and object-based approaches. This technique enables the tool to achieve an accuracy exceeding 85% with 15 classes.

The research contributes novel solutions for handling compositional data and delivers a high-performing mapping tool for coral reef ecosystems, aiding in environmental management and conservation efforts.

Résumé

Modèles d'apprentissage automatique pour la cartographie de récifs coralliens par imagerie satellite

La crise actuelle du changement climatique nécessite le développement de méthodes efficaces pour surveiller et cartographier l'environnement et les espèces afin d'assurer leur préservation. Cette thèse explore l'application d'algorithmes d'apprentissage automatique pour cartographier efficacement les récifs coralliens à partir d'images satellites multispectrales. Le lagon de Maupiti en Polynésie française sert d'étude de cas. Ce travail de recherche a conduit à la production d'un outil automatisé capable de générer des cartes de récifs coralliens à partir d'images satellites. De plus, cet outil peut être adapté pour cartographier d'autres écosystèmes, tels que des forêts ou des calottes glaciaires, à condition que le modèle soit ré-entraîné avec des données pertinentes.

Tout d'abord, une analyse de la littérature scientifique des dernières années examine les méthodes et les tendances actuelles en matière d'utilisation d'algorithmes d'apprentissage automatique pour la cartographie des récifs coralliens. Ensuite, les tentatives de développement de l'outil nous ont conduits à nous confronter au cas particulier des données compositionnelles, qui sont des données contenant des informations relatives et se situant dans un espace mathématique connu sous le nom de simplexe. Des adaptations des méthodes conventionnelles sont nécessaires pour répondre aux caractéristiques spécifiques de cet espace.

Dans un premier temps, en réponse aux jeux de données déséquilibrés, une technique de suréchantillonnage est développée spécifiquement pour les données compositionnelles. En outre, un modèle autorégressif spatial basé sur la distribution de Dirichlet est formulé pour tenir compte des effets spatiaux qui peuvent survenir dans le processus de cartographie automatique.

Enfin, nous présentons la mise en oeuvre de notre outil de cartographie final. Pour atteindre l'objectif souhaité, un processus de classification en deux étapes est mis en place, combinant des approches basées sur les pixels et sur les objets. Cette technique permet à notre outil d'atteindre une précision supérieure à 85% avec 15 classes.

Ce travail de recherche apporte de nouvelles solutions pour le traitement des données compositionnelles et fournit un outil de cartographie performant pour les écosystèmes de récifs coralliens, pouvant contribuer à la gestion de l'environnement et aux efforts de conservation.

Contents

CHAPTER 1

General introduction

1.1 The Anthropocene and the collapse of biodiversity

Over the past five millennia, one-third of the Earth's forests have been lost [349]. Within the last five decades, two-thirds of the vertebrate population has vanished [10]. Every decade, 9% of the insect population disappears [425]. These three figures sum up the situation we are facing.

This era of destruction brought by human activities is called the Anthropocene [455]. The solutions exist and have been known for at least 35 years, when the first IPCC report was published [86]. However, they still have not been implemented.

The collapse of biodiversity is observed worldwide, even within supposedly protected areas [169], [234]. Much of this decline can be attributed to the adverse impacts of habitat loss and direct human activity, rather than being primarily driven by climate change at this point in time. This alarming trend spans across diverse landscapes, including deserts [110], [200], forests [36], [39], [130], [247], grasslands [131], [330], [402], and oceans [203].

In the past century, the average temperature on the ocean surface has increased by 0.88°C [123]. Marine heatwaves, already longer, more frequent and intense than a century ago, will become 4 times more frequent in a few decades, under the best scenario [123]. Since the mid-20th century, it is known [347] that oceans absorb approximately one-quarter of the carbon dioxide produced by human activities [358], as described by the chemical equation:

$$CO_2 + H_2O \leftrightarrow H^+ + HCO_3^- \leftrightarrow 2H^+ + CO_3^{2-}$$

According to literature [308], an increase in the concentration of hydrogen ions (H^+) resulting from this process leads to a decrease in pH, causing ocean acidification, impacting the lives of marine organisms [228].

Furthermore, some studies suggest that CO2 absorption will lead to undersaturation of aragonite in the Southern Ocean by 2050 [309], posing a threat to pteropods with aragonite shells. Over a longer timescale of thousands of years, oceans are projected to absorb over 90% of total anthropogenic CO2 emissions [25]. This would result in the reversal of sedimentation rates in the oceans, with CaCO3 dissolving faster than it is produced.

On top of that, climate change also leads to a decline in the oxygen concentration in the oceans [54], significantly affecting the entire aquatic ecosystem. Given that oxygen is essential for the survival of all marine organisms, these alterations have far-reaching consequences for marine life.

Aside from the challenges posed by climate change, marine life faces additional threats. A major concern for ocean health is overfishing [135], [318], which is pushing marine ecosystems

to critical levels and endangering the conservation status of various species, driving some of them close to extinction [109], [359], [449]. In Europe, more than half of the fisheries stocks are outside of safe biological limits [127].

Among the threaten marine ecosystems, the most complex one and the more at risk is probably coral reefs.

1.2 Coral reefs

Coral reefs are highly diverse and intricate ecosystems, hosting a multitude of interacting species, some of which have roles that are not yet fully understood [444]. Reefs comprise over 850 coral-associated invertebrate species [142], many of which are interdependent [337] and play crucial roles in marine ecosystems, extending beyond the boundaries of the reefs themselves [360].

At the close of the 20th century, global reef coverage was estimated to be approximately 255,000 km^2, which is roughly equivalent to the size of the United Kingdom [401]. Despite occupying less than 0.01% of the total ocean surface [96], reefs were estimated to support 5% of the global biota by the late 1990s [346] and harbor 25% of all marine species [350].

Moreover, coral reefs provide an array of valuable services and products, amounting to a staggering annual value of 172 billion USD per km^2 [273], equivalent to around 2000 times the United States' GDP. They provide coastal protection from floods and storms, and a disappearance of reefs would result in an increase in anticipated damages, with flood-related impacts doubling and storm-related damages tripling [40].

Unfortunately, coral populations are currently in decline due to rising global temperatures [348], exacerbating the vulnerability of these ecosystems as higher sea surface temperatures are strongly associated with coral bleaching [339]. A study spanning the past two decades has demonstrated that areas experiencing more frequent and intense thermal-stress anomalies exhibit greater instances of coral bleaching [405]. The link between ocean warming and coral bleaching was already established three decades ago [145]. The susceptibility of the highly vulnerable *Acropora* corals to bleaching is evidenced by events such as the 2000 thermal stress incident in the Saipan lagoon, which resulted in the mortality of approximately 40% of the affected corals [191], [325].

While heat stress is a significant cause of bleaching, other factors can also induce bleaching, including light stress [237], herbicide exposure [295], and oxygen depletion [105].

Severe bleaching events have been on the rise over the past four decades [196], and numerous models predict that by 2100, over 95% of reefs will experience severe bleaching events at least twice per decade [253]. Additionally, corals have limited capacity to recover fully from repeated bleaching events [197], rendering them less resilient and more susceptible to the effects of ocean warming [370].

It is important to note that bleaching does not imply immediate coral death; however, it does have detrimental effects such as atrophy, necrosis, increased mortality [146], reduced recovery from diseases [310], and loss of architectural complexity [331], which importance will be discussed later. Furthermore, coral taxa exhibit varying responses to bleaching events [198], [270], as evidenced by the shift in the proportions of soft and hard corals in the Great Barrier Reef between 2012 and 2017 due to bleaching coupled with tropical cyclones [429].

In 2009, a study [103] examined 69 coral reefs in the Great Barrier Reef and indicated that increasing temperature stress was impacting the ability of corals to deposit calcium carbonate. The study revealed a significant decline of 14.2% in the calcification rate between 1990 and 2009, a rate that had not been observed in the past 400 years. However, it is important to note that the reduction in calcification rate is primarily observed when stress levels exceed a temperature threshold of 3°C and when pCO_2 levels surpass 700ppm [225]. Under these conditions, coral calcification can decrease by approximately 20%.

Ocean acidification and the absorption of carbon dioxide reduces the concentration of carbonate ions (CO_3^{2-}), which in turn hampers coral calcification [221]. A model developed by [285] and based on a study of five reef sites predicts a decline in skeletal density of *Porites* corals between 15% and 25% by the end of the 21st century due to ocean acidification alone. However, studies such as [9] demonstrate that the effects of ocean acidification on coral reef calcification can be reversible. In this study, reef calcification increased by 7% within a week when the water chemistry was restored to pre-industrial conditions through alkalinity enrichment. Similar findings are observed in [372], which investigates the site of Heron Island (Australia), where human activities in the mid-20th century led to a decrease in water levels, exposing corals to air during low tide. The study shows that between 1972 and 2011, corals experienced growth and recovery due to the construction of walls that raised water levels. Other examples from the literature indicate that coral reefs can be resilient and recover relatively quickly as long as the stress factors are not too severe [310], [371], [428].

Nevertheless, despite the potential for coral recovery from various stress factors, it has been documented that their recovery rate has declined by 84% in the Great Barrier Reef over the past 30 years [310].

The architectural complexity of coral reefs has been identified as a critical indicator of coral health [161]. Beyond its role as an indicator, it plays a vital role in shaping fish assemblages and regulating factors such as predation and competition through the provision of diverse habitats. This complexity further enhances the dissipation of wave energy and momentum compared to the simpler sandy beach environments [102], [396]. Diminished architectural complexity has been shown to amplify the coastline's susceptibility to wave impacts [72]. Conversely, waves act as conveyors of nutrients, facilitating mixing and oxygenation within the ecosystem. The synergy among wave dynamics, reef biology, and architectural complexity underscores the significance of the latter in monitoring reef evolution and developing hydrodynamics models of habitats [394], [399]. Lastly, the biological significance of architectural complexity extends further as it also has an impact on coral reef fishes. While many of these species display resilience to coral bleaching, their tolerance dwindles when the integrity of the reef's structure is compromised [331].

In addition to bleaching, other factors can impact the architecture of coral reefs, such as ocean acidification [221] and disturbances, which can have lasting consequences on reef structure even after apparent recovery [48]. A study conducted in 2009, focusing on 200 Caribbean reefs, highlighted a global decline in structural complexity irrespective of water depth [11].

Human activities, particularly explosive fishing, have been identified as significant contributors to the decline of coral reefs [28]. A study conducted by [304] on Suranti Island in Indonesia examined data from 1972 to 2013 and revealed a substantial reduction in the proportion of live coral coverage, exceeding 75%. This decline was primarily attributed

to explosive fishing. The researchers observed a significant increase in rubble-covered areas, which expanded from nonexistent in the 1980s to encompassing 330 hectares in 2013, equivalent to half of the study site's total area. Furthermore, reefs face threats from plastic waste, which increases the risk of disease by 40 times [233], with complex reefs being eight times more susceptible to the effects of plastic waste. This number is even more alarming knowing that approximately 9.5 million tons of plastic enter the oceans every year [52], [206].

Given the profound implications of climate change on the oceans, particularly coral reefs, this thesis is driven by the imperative of monitoring their dynamic evolution and creating precise habitat maps.

1.3 Monitoring the evolution of coral reefs

Monitoring the evolution of environmental features or species poses a crucial challenge [255] as it serves two key purposes: firstly, enabling the prediction of species' evolution [422], and more significantly, facilitating interventions to positively influence their changes and support their recovery or enhanced conservation [62], [100]. In the context of coral reefs, mapping plays a pivotal role in their monitoring, encompassing various aspects such as identifying coral zones and distinguishing healthy coral from deteriorated or bleached coral [178], [321]. The collection of data serves as the foundation for conducting these mappings. Numerous data-gathering techniques can be employed, including but not limited to: divers capturing underwater photos or videos, boats equipped with underwater imaging or lidar systems, drones, and satellite imagery [154]. Each of these techniques possesses its own advantages and disadvantages, making them suitable for specific purposes and scenarios.

Diver-based data collection provides the advantage of capturing high-resolution images or videos in close proximity to the coral reef, enabling detailed observations and valuable qualitative information. However, this approach is labor-intensive, time-consuming, and limited in coverage due to the restricted range of divers. On the other hand, boats and drones offer the benefit of wider coverage compared to divers. However, both techniques require some level of human presence on-site. For areas highly inhabited by humans and scientists, such as the Great Barrier Reef in Australia, having regular human presence for data collection is feasible. Scientists can easily visit the site and gather data. However, for remote sites with limited human population, such as some Pacific islands with only a few hundred inhabitants, maintaining a regular data collection schedule becomes more challenging. Scientists may not have the resources to travel to these locations every year or even every few years.

Satellite imagery presents a distinct advantage in terms of wide coverage area and the ability to capture data on a regular basis [337]. This capability enables long-term monitoring without the need for an onsite presence. By leveraging satellite technology, scientists can obtain valuable data from remote locations, including islands with limited human population. However, satellite imagery does have limitations that need to be considered. The spatial resolution of satellite data is typically coarser compared to other data collection techniques. Fine-scale details such as distinguishing between different coral species or achieving sub-meter resolution may not be feasible with satellite imagery alone [179]. However, despite these limitations, satellite imagery still serves valuable purposes in coral reef monitoring and research. For instance, it can be utilized effectively in studying the hydrodynamics of waves

on the shoreline, where the focus is on larger-scale patterns and trends rather than fine-scale details. The broad coverage and regular data availability of satellite imagery also make it well-suited for analyzing the global coral coverage and assessing large-scale changes over time.

While other challenges such as cloud cover and water turbidity can affect the quality of satellite imagery, various preprocessing techniques exist to address these issues. Some preprocessing methods, such as atmospheric correction and cloud masking, can help mitigate the impact of these factors and improve the usability of the imagery for mapping and analysis purposes [174], [245], [462].

1.4 Machine learning for satellite images

1.4.1 Context

Although satellite images are readily available over wide areas and on a regular basis, their analysis by human experts remains a time-consuming task. For instance, the size of the Great Barrier Reef is more than 348,000km^2 [421]. Even with comprehensive data available for this reef, the manual mapping of such an extensive area would be impractical for an individual or even a group.

Machine learning represent a possible solution to this problem. It is widely used nowadays across a large range of fields [210], providing automation for tasks and computations that would otherwise be arduous for humans to handle. Machine learning algorithms can be broadly classified into two main categories [287], [365], [369]:

- Supervised learning. In this category, the computer is provided with data that consists of both input and corresponding output information. By analyzing the input data, the model aims to identify a formula or relationship that can generate output data as accurately as possible.
- Unsupervised learning. Unlike supervised learning, the computer is only given access to the input data without any corresponding output information. The computer then tries to find patterns or structures inherent within the input data.

Machine learning models, particularly through the use of supervised learning, offer an effective approach for mapping tasks based on satellite imagery [383], [407]. In this context, the satellite image serves as the input, containing various color channels and visual information, while the desired output is the map that we aim to generate. **Because the field of machine learning encompasses a wide range of models, further insights on the most efficient methods on this specific case of satellite imagery and coral reefs is needed.**

1.4.2 Artificial Neural Networks and Random Forests

Among the various methods available, two standout approaches are Artificial Neural Networks (NN) and Random Forests (RF). These methods hold a prominent position in the field due to their widespread utilization and consistent delivery of superior outcomes.

Neural networks have become increasingly used over the past years [369] and are inspired by the structure and functioning of biological neural networks in the human brain. An

artificial neural network consist of several layers of interconnected nodes, or neurons. Each neuron takes an input, performs calculations, and returns an output. The combination of all the neurons and layers allows to produce complex operations and behaviour. The weights of the NN, i.e. the values of the parameters with which the computations are performed, are learnt by minimizing the error between the predicted and the expected output.

On the other hand, RF is an ensemble of decision trees. A decision tree takes on the form of a flowchart, wherein each internal node corresponds to a specific feature or attribute, each branch represents a decision rule, and each leaf node represents an outcome. The goal of the decision tree is to recursively split the data into smaller and more homogeneous subsets until they are sufficiently pure to make accurate predictions at the leaf nodes.

While Neural Networks have gained immense popularity in recent times, we will thoroughly compare and evaluate both methods to ultimately adopt Random Forests in the final workflow.

1.4.3 Pixel-based and object-based

When using machine learning algorithms with images for mapping purposes, two primary strategies can be distinguished: pixel-based and object-based approaches [410]. In the first one, each pixel is classified independently without considering its surrounding context. This strategy can be broadened to include the kernel approach, wherein sets of pixels are evaluated within a kernel filter or moving window — it is worth noting that this can be regarded as a distinct method in itself. In this thesis, all the classified pixels are considered pure, i.e. each pixel is associated to a single class. On the other hand, the object-based approach involves constructing and classifying objects composed of multiple pixels. To achieve this, image segmentation techniques are commonly employed. Image segmentation aims to partition an image into distinct and meaningful regions or objects. This process typically involves grouping pixels based on their color, texture, or other characteristics, as well as detecting edges within the image [214], [391].

Segmentation techniques provide objects, called segments, containing several pixels. Consequently, within each segment, various values can be computed based on the pixel attributes. For instance, statistical moments such as the mean or other descriptive statistics can be calculated for the reflectance values of the pixels across each spectral band. Besides, it is important to note that when we have an expert-based map available for the segmented image, the segment boundaries may not align precisely with the expert's delineations. In fact, it is rare for the segment boundaries to perfectly match those defined by the expert. As a result, each segment may contain a mixture of pixels belonging to different classes, rather than exclusively representing a single class. Hence, if we denote J the number of existing classes, each segment i can be assigned a vector $y_i \in [0,1]^J$ representing the proportions or probabilities of the pixels within the segment belonging to each respective class. Each element of the vector y_i represents the proportion of pixels in segment i that belong to a specific class, ranging from 0 (no pixels) to 1 (all pixels). By using such a vector representation, we can capture the distribution and class membership information for each segment, allowing for further analysis and classification tasks based on the pixel proportions within the segment. Such a vector is called a **compositional data**, having distinct properties and characteristics due to its specific shape and constraints [5]. Figure 1.1 gives an example of such vectors, with three classes and four segments.

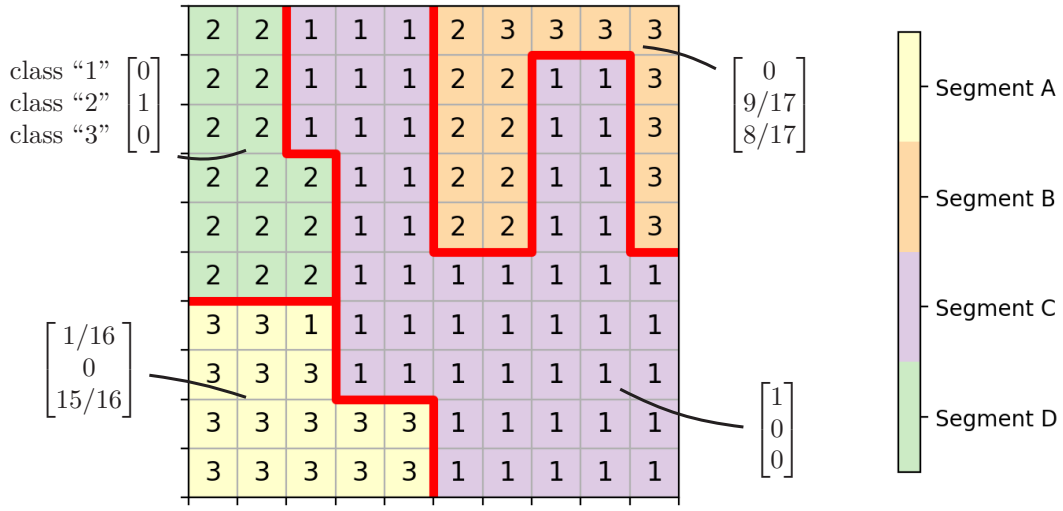


Figure 1.1: Example of a segmentation on an image of size 10×10 pixels, with four segments (A, B, C, D) and three classes (1, 2, 3). In each segment, we compute the ratio of pixels belonging to each class.

1.4.4 Compositional data

Mathematically, the compositional data vectors lie within a space known as a simplex and denoted S^{J-1},

$$S^{J-1} = \left\{ y = (y_1, y_2, ..., y_J) \mid \forall i \in \{1, 2, \ldots, J\}, y_i \geq 0; \sum_{i=1}^{J} y_i = 1 \right\}. \tag{1.1}$$

A simplex can be considered as an extension of the concept of a triangle to higher dimensions. In the case of the simplex space S^{J-1}, a point belonging to it is a vector of size J that can be described using only $J-1$ coordinates, as the last coordinate can be derived from the others.

The vectors y_i represent the response variables (or labels) that we aim to predict using machine learning models. However, due to the constraints imposed by the simplex space, we need to adapt existing machine learning methods. While it is possible to consider the majority class of each vector y_i (i.e., the value $\text{argmax}_j(y_{ij})$) and apply conventional classification methods, this approach leads to a loss of information and is not desirable. Therefore, a significant portion of this thesis involves the adaptation of methods to effectively handle compositional data labels.

One natural distribution to handle compositional data is the Dirichlet distribution. It is parameterized by a vector $\alpha \in \mathbb{R}^J$ such that for each $j \in [1, \ldots, J]$, $\alpha_j > 0$. Its probability density function is given by

$$f(y|\alpha) = \frac{\Gamma(\sum_{j=1}^{J} \alpha_j)}{\prod_{j=1}^{J} \Gamma(\alpha_j)} \prod_{j=1}^{J} y_j^{\alpha_j - 1}, \tag{1.2}$$

where Γ is the gamma function.

The utilization of the Dirichlet distribution holds particular significance in this thesis as we expect the labels y_i of each segment to follow such a distribution. Consequently, we can develop a maximum likelihood estimator to map each segment to its corresponding expected compositional data label. Additionally, since neighboring segments might induce a spatial

effect and impact each other, we may consider whether **incorporating a spatial lag term into the model can enhance its performance**.

1.5 Case-study: Maupiti lagoon

The main focus of this thesis revolves around the development of the tool based on data from Maupiti Island in French Polynesia. Maupiti Island is part of the Society Islands archipelago, and its lagoon spans approximately 27km^2. The ground-truth data used for this study is derived from expert-based mapping, which involved multiple field observation campaigns [394]. Figure 1.2 depicts a satellite image of the island along with the expert mapping. What sets this dataset apart is that the ground-truth map covers the entire area of Maupiti, whereas most studies only have access to a limited number of ground-truth points. This unique characteristic allows us to employ techniques that are not feasible or less accurate when working with sparse ground-truth data, such as object-based classification.

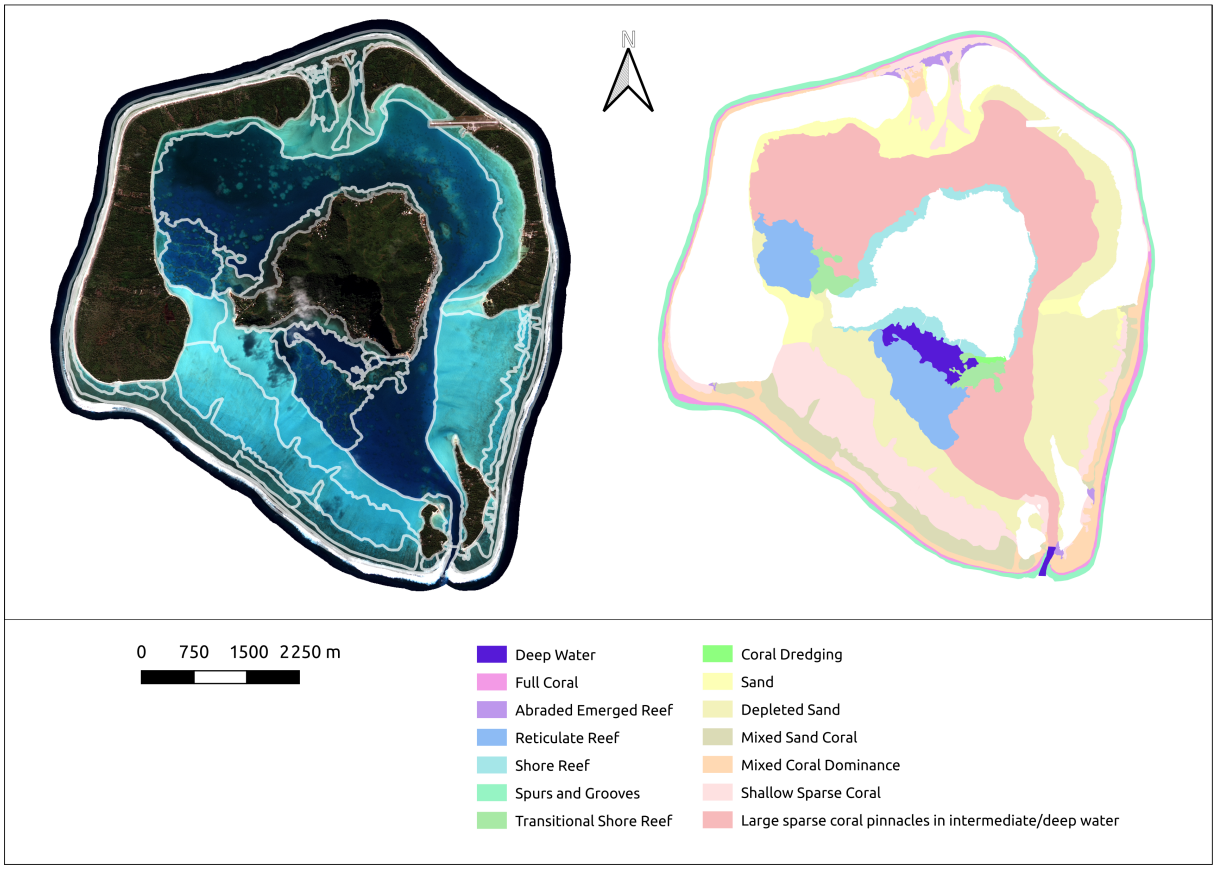


Figure 1.2: Pleiades-2 satellite image of Maupiti island, with the delineation of the zones (left) and the expert-based map of the site (right).

Several satellite images are available for the study site. The primary dataset used extensively in this research is the freely accessible Sentinel-2 images, which have a resolution of 10m. These images are particularly suitable for our purposes due to their lower resolution, which results in a smaller dataset to work with, facilitating the development of the workflow in a more efficient and expedited manner. Additionally, we acquired a higher-resolution image from the Pleiades satellite, with a resolution of 2m, enabling more precise mapping at

the cost of increased processing time. It is important to note that the workflow developed in this thesis is specific to the satellite source it was trained on, and its application to images from different satellite sources will be discussed in detail during the thesis discussion.

In comparison to other regions heavily impacted by climate change and human activity, such as the Great Barrier Reef, the coral reefs of Maupiti Island, like many other islands in French Polynesia, are relatively healthy [149]. This distinction must be considered when applying the developed tool to areas with different conditions, as the classification may encounter challenges due to these variations.

1.6 Objectives of the thesis

Mapping coral reefs worldwide in a scalable and time-efficient manner is of utmost importance. To achieve this objective, a combination of machine learning models and satellite imagery emerges as a promising solution. Within this framework, the general objective of this thesis is to assess the performances of various machine learning models in mapping coral reefs from satellite images using full expert-based maps as ground-truth data. Our specific focus lies in understanding how the identified models are suited to address the compositionality that arises within the labels during image segmentation, and what are the advantages of such compositional labels.

To address these inquiries, this thesis is structured around four key objectives:

- Literature Review: The first step involves conducting a comprehensive literature review to gain insights into the current state of coral reef mapping using multispectral imagery. Chapter 2 serves as this extensive review, focusing specifically on papers published between 2018 and 2020.

- Data Imbalance Correction: Compositional data, particularly in the context of coral reef mapping, often exhibit class imbalance, with varying numbers of elements in each class. Chapter 3 introduces an oversampling technique tailored to rectify this data imbalance effectively.

- Spatial Considerations: Given the inherent spatial nature of the mapping process, it is crucial to consider spatial dependencies. Chapter 4 introduces the integration of spatial components into a Dirichlet regression model to better capture these spatial relationships.

- Tool Development: The culmination of this research is the development of a robust mapping tool for coral reefs derived from multispectral satellite imagery, trained with a full ground-truth map, using both the spectral and spatial dimensions of the image. This final workflow is presented in Chapter 5.

Finally, the conclusion of this thesis encompasses the response to the central research questions, encapsulating the scientific contributions made throughout its course. The concluding section also delves into potential future avenues and areas that require further exploration.

CHAPTER 2

Mapping of Coral Reefs with Multispectral Satellites: A Review of Recent Papers

The work in this chapter is a paper published in Remote Sensing: T. Nguyen, B. Liquet, K. Mengersen, and D. Sous. Mapping of coral reefs with multispectral satellites: a review of recent papers. *Remote Sensing*, 13(21):4470, 2021 [298].

Synopsis

This chapter provides a comprehensive review of the literature on the automatic mapping of coral reefs using multispectral satellite images. Over the past few years, there has been a significant surge in research papers addressing this topic. This literature review particularly emphasizes the papers published between 2018 and 2020, considering the recent influx of research on coral reef mapping.

Within this review, we explore various aspects, including the quality and cost analysis of different satellite image sources, the efficacy of diverse satellite imagery preprocessing techniques, a comparison between object-based and pixel-based classification methods, and the evaluation of different machine learning models.

By delving into these areas, this literature review serves as a gateway to identifying the most suitable techniques and models to be employed throughout the thesis.

Abstract

Coral reefs are an essential source of marine biodiversity, but they are declining at an alarming rate under the combined effects of global change and human pressure. A precise mapping of coral reef habitat with high spatial and time resolutions has become a necessary step for monitoring their health and evolution. This mapping can be achieved remotely thanks to satellite imagery coupled with machine-learning algorithms. In this paper, we review the different satellites used in recent literature, as well as the most common and efficient machine-learning methods. To account for the recent explosion of published research on coral reel mapping, we especially focus on the papers published between 2018 and 2020. Our review study indicates that object-based methods provide more accurate results than pixel-based ones, and that the most accurate methods are Support Vector Machine and Random Forest. We emphasize that the satellites with the highest spatial resolution provide the best images for benthic habitat mapping. We also highlight that preprocessing steps (water column correction, sunglint removal, etc.) and additional inputs (bathymetry data, aerial photographs, etc.) can significantly improve the mapping accuracy.

2.1 Introduction

Coral reefs are complex ecosystems, home to many interdependent species [142] whose roles and interactions in the reef functioning are still not fully understood [444]. By the end of the 20th century, reefs were estimated to cover a global area of 255,000 km^2 [401], which is roughly the size of the United Kingdom. Although this number represents less than 0.01% of the total surface of the oceans [96], reefs were estimated to be home to 5% of the global biota at the end of the 1990s [346] and to 25% of all marine species [350]. Furthermore, each year reefs provide services and products worth the equivalent of 172 billion US$ per km^2 [273], thus "producing" a total equivalent of 2000 times the United States GDP.

Despite their importance, coral populations are collapsing due to several factors mainly driven by global climate change and human activity. One of the main threats is the rise of global temperatures [348]. Increasing sea surface temperature is strongly correlated with coral bleaching [339], which tends to be enhanced by the intensity and the frequency of thermal-stress anomalies [405]. Bleaching does not mean that the corals are dead, but it leads to a series of adverse consequences: atrophy, necrosis, increase of the death rate [146], less efficient recovery from disease [310] and loss of architectural complexity [331]. Repeated bleaching events are even more damaging, impairing the coral colony recovery [197] and making them less resistant and more vulnerable to ocean warming [370]. These cumulative impacts are particularly alarming when considering the increasing frequency of severe bleaching events in the last 40 years [196] and predictions that by 2100, more than 95% of reefs will experience severe bleaching events at least twice a decade [253].

The worldwide decline of coral reefs has prompted an unprecedented research effort, reflected by the exponential growth of scientific articles dedicated to coral reefs (see Figure 2.1). A key prospect faced by the scientific community is the development of open and robust monitoring tools to survey the reef distribution on a global scale for the next decades. Mapping benthic reef habitats is crucially important for tracking their time and space evolution, with direct outcomes for reef geometry and health surveys [161], for developing

numerical models of circulation and wave agitation in reef-lagoon systems [394], [399] and for socio-economic and environmental management policies [175].

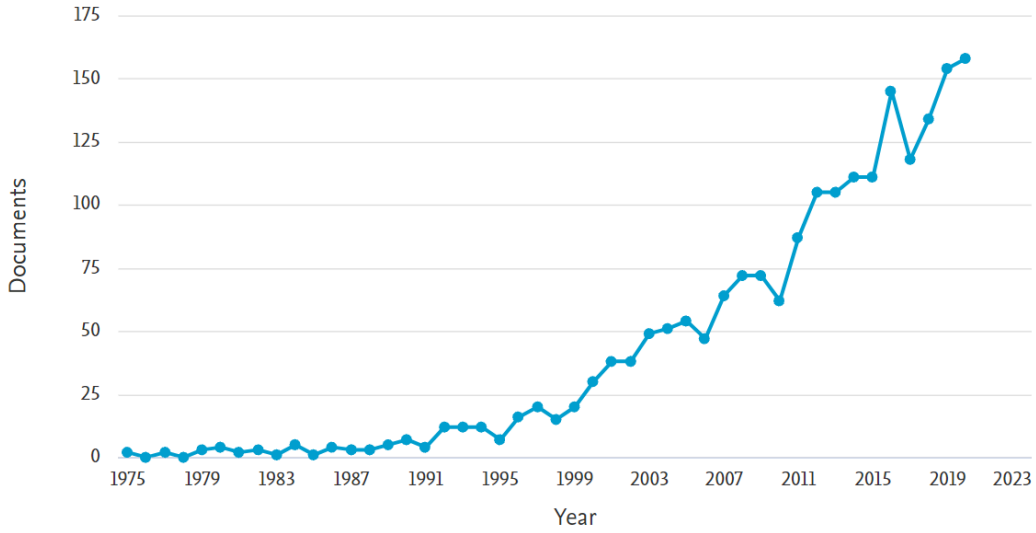


Figure 2.1: Number of documents tagging "coral mapping" or "coral remote sensing" in the Scopus database over the last 50 years.

A powerful tool to survey coral reefs is coral mapping or coral classification. This involves using raw input data of a coral site, such as videos or images, extracting the characteristics of the ground and classifying the elements as coral, sand, seagrass, etc. To perform this mapping, there are two possibilities: manually extracting the characteristics, which is a highly accurate method but tedious and time consuming, or training machine-learning algorithms to easily do it in a short time but with a higher chance of misclassification. In this article, the terms "coral mapping" and "coral classification" will both refer to the same meaning being the "automatic machine-learning mapping" if not otherwise stated.

Coral mapping can be accurately achieved from underwater images, as done in most papers published in 2020 [147], [254], [264], [268], [280], [281], [317], [344], [412], [430], [452]. However, a major drawback of underwater images is that they are difficult to acquire at a satisfying time resolution for most remote places, thus making it unfeasible to have a worldwide global map with this kind of data. One solution is to use data from satellite imagery.

Aiming to help the ongoing and future efforts for coral mapping at the planetary scale, this paper will mainly focus on multispectral satellite images for coral classification and will mostly omit other sources of data. The main goal of this paper is to highlight the current most efficient methods and satellites to map coral reef. As depicted in Figure 2.1, there are twice as many papers published in the past two years than there were ten years ago. Furthermore, as described later, the resolution of satellites is quickly improving, and with it the accuracy of coral maps. This is also true for machine-learning methods and image processing. Finally, substantive reviews of work related to coral mapping are only available to 2017 [170], [337]. For these reasons, we decided to narrow our analysis to papers published since 2018. Between 2018 and 2020, 446 documents tagging "coral mapping" or "coral remote sensing" have been published (Figure 2.1). However, most of these papers do not fit within the scope of our study: they are for instance treating tidal flats, biodiversity

problems, chemical composition of the water, bathymetry retrieval, and so on. Thus, out of these 446, only 75 deal with coral classification or coral mapping problems. The data sources used in these papers are summarized in Figure 2.2. Within these 75 studies, a subset of 37 papers that deal with satellite data (25 with satellite data only) will be specifically included in the present study.

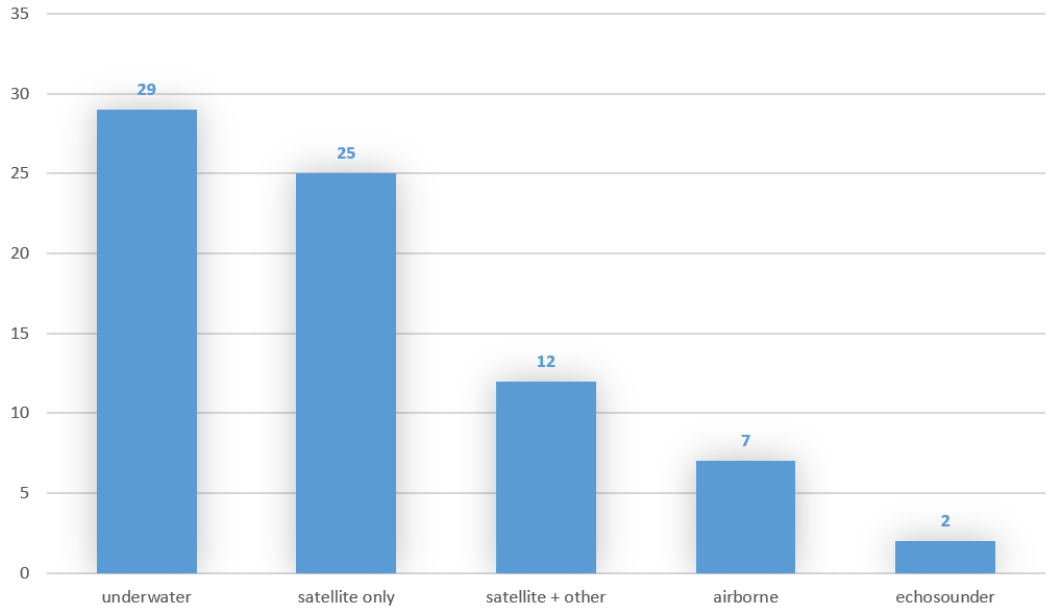


Figure 2.2: Bar plot presenting the data sources of 75 different papers from 2018 to 2020 studying corals classification or corals mapping.

Used in almost 50% of the papers, satellite imagery is recommended by the Coral Reef Expert Group for habitat mapping and change detection on a broad scale [150]. It allows benthic habitat to be mapped more precisely than via local environmental knowledge [377] on a global scale, at frequent intervals and with an affordable price.

This review is divided into four parts. First, the different multispectral satellites are presented, and their performance compared. Following this is a review of the preprocessing steps that are often needed for analysis. The third part provides an overview of the most common automatic methods for mapping and classification based on satellite data. Finally, the paper will introduce some other technologies improving coral mapping.

2.2 Satellite Imagery

2.2.1 Spatial and Spectral Resolutions

When trying to classify benthic habitat, two conflicting parameters are generally put in balance for choosing the satellite image source: the spatial resolution (the surface represented by a pixel) and the spectral resolution. The latter generally refers to the number of available spectral bands, i.e., the precision of the wavelength detection by the sensor. The former parameter has a straightforward effect: a higher spatial resolution will allow a finer habitat mapping but will require a higher computational effort. The primary effect of the spectral resolution is that model accuracy generally increases with the number of visible bands [51], [90], [266] and the inclusion of infrared bands [91]. Although no clear definition exists,

a distinction is generally made in terms of spectral resolution between multispectral and hyperspectral satellites. The former sensors produce images on a small number of bands, typically less than 20 or 30 channels, while hyperspectral sensors provide imagery data on a much larger number of narrow bands, up to several hundred—for instance NASA's Hyperion imager with 220 channels. Most of the time, multispectral and hyperspectral sensors have an additional panchromatic band (capturing the wavelengths visible to the human eye) with a slightly higher spatial resolution than the other bands.

A major drawback of hyperspectral satellites is that the best achievable resolution is generally several tens of meters and can be up to 1km for some of them [178], while most multispectral sensors have a resolution better than 4m. A high spectral resolution coupled with a low-spatial-resolution result in a problem known as "spectral unmixing", which is the process of decomposing a given mixed pixel into its component elements and their respective proportions. Some existing algorithms can tackle this issue with a high level of accuracy [49], [165], [433]. When unmixing pixels, algorithms may face errors due to the heterogeneity of seabed reflectance, disturbing the radiance with the light scattered on the neighboring elements [164]. This process, called the adjacency effect, has negative effects on the accuracy of remote sensing [59] and can modify the radiance by up to 26% depending on turbidity and water depth [73].

In this review, we purposefully omitted the hyperspectral sensors to focus on multispectral satellite sensors, since only the latter have a spatial resolution fine enough to map coral colonies. Moreover, in our case where we are studying how to create high-resolution maps of coral presence, multispectral satellites are more efficient, i.e., they provide more accurate results [150]. In the following parts, unless otherwise stated, the spatial resolution will be referred to as "resolution".

2.2.2 Satellite Data

We found 14 different satellites appearing in benthic habitat mapping studies, and gathered in Table 2.1 their main characteristics, in particular their spectral bands, spatial resolution, revisit time and pricing. The Landsat satellites prior to Landsat 6 do not appear in the table because they are almost universally not used in recent studies, the Landsat 5 being deactivated in 2013.

The most commonly used multispectral satellite images are from NASA's Landsat program [342]. The program relies on several satellites, of which Landsat 8 OLI, Landsat 7 ETM+, Landsat 6 ETM and Landsat TM have been used for benthic habitat mapping: [30], [77], [132], [133], [204] (OLI), [28], [136], [304] (OLI, ETM+, TM), [18], [19], [115], [314] (ETM+). The standard revisit time for Landsat satellites is 16 days. However, Landsat-7 and Landsat-8 are offset so that their combined revisit time is 8 days. The density and accuracy of the Landsat images thus make them viable to use for ecological analyzes [217].

Sentinel-2 [177], a European Space Agency's satellite, can be compared to Landsat satellites in terms of spatial and spectral resolution. Sentinel-2 was initially designed for land monitoring [271] but has been used for monitoring oceans (and more specifically coral reefs bleaching) and mapping benthic habitat [55], [202], [230], [240], [357], [378], [445]. Specific spectral bands of Sentinel-2, such as SWIR-cirrus and water vapor bands, are especially useful for cloud detection and removal algorithms [33], [111], [361], [374], [386]. One major

advantage of Landsat and Sentinel-2 satellites is that their data are open access. However, these satellites are defined as "low-resolution", with a resolution of tens of meters which may be a significant weakness when trying to map and to classify the fine and complex distribution of coral reef colonies.

With a typical spatial resolution of several meters, medium-resolution satellites are more accurate than the aforementioned satellites. Well-known medium-resolution satellites are SPOT-6 [387], [406] and RapidEye [88], [141], [307], with respectively 4 and 5 bands. A major strength of RapidEye is that the image data are produced by a constellation of five identical satellites, thus providing images at a high frequency (global revisit time of one day). Note however that up to now, RapidEye has not been found in recent literature for coral mapping. The principle of using multiple similar satellites is also found with the PlanetScope constellation, composed of 130 Planet Dove satellites. Their total revisit time is less than one day, and they can be found in several recent coral mapping studies [29], [242], [353], [439], [457].

Table 2.1: Comparison of some characteristics of the most common multispectral satellites. Excepted for PlanetScope and RapidEye, all the satellites contain a panchromatic band which does not appear in the column "Spectral bands". Image pricings have been recovered from the website wwww.apollomapping.com accessed on February 2021

Satellite Name	**Spectral Bands**	**Resolution (at Nadir)**	**Revisit Time**	**Pricing**
Landsat-6 ETM	4 VNIR 2 SWIR 1 thermal infrared	15 m panchromatic 30 m VNIR and SWIR 120 m thermal	16 days	Free
Landsat-7 ETM+	4 VNIR 2 SWIR 1 thermal infrared	15 m panchromatic 30 m VNIR and SWIR 60 m thermal	16 days	Free
Landsat-8 OLI	4 VNIR 3 SWIR 1 deep blue	15 m panchromatic 30 m VNIR and SWIR 30 m deep blue	16 days	Free
Sentinel-2	4 VNIR 6 red edge and SWIR 3 atmospheric	10 m VNIR 20 m red edge and SWIR 60 m atmospheric	10 days	Free
PlanetScope	$\emptyset$ panchromatic 4 VNIR	$\emptyset$ panchromatic 3.7 m multispectral	<1 day	\$1.8 /km^2
RapidEye (five satellites)	$\emptyset$ panchromatic 5 VNIR	$\emptyset$ panchromatic 5 m multispectral	1 day	\$1.28 /km^2
SPOT-6	4 bands: blue, green, red, near-infrared	1.5 m panchromatic 6 m multispectral	1–3 days	\$4.75 /km^2
GaoFen-2	4 bands: blue, green, red, near-infrared	0.81 m panchromatic 3.24 m multispectral	5 days	\$4.5 /km^2
GeoEye-1	4 bands: blue, green, red, near-infrared	0.41 m panchromatic 1.65 m multispectral	2–8 days	\$17.5 /km^2
IKONOS-2	4 bands: blue, green, red, near-infrared	0.82 m panchromatic 3.2 m multispectral	3–5 days	\$10 /km^2
Pleiades-1	4 bands: blue, green, red, near-infrared	0.7 m panchromatic 2.8 m multispectral	1–5 days	\$12.5 /km^2
Quickbird-2	4 bands: blue, green, red, near-infrared	0.61 m panchromatic 2.4 m multispectral	2–5 days	\$17.5 /km^2
WorldView-2	8 VNIR	0.46 m panchromatic 1.84 m multispectral	1.1–3.7 days	\$17.5 /km^2
WorldView-3	8 VNIR 8 SWIR 12 CAVIS	0.31 m panchromatic 1.24 m VNIR 3.7 m SWIR 30 m CAVIS	1–4.5 days	\$22.5 /km^2

Finally, high-resolution sensors are defined as those with a few meters resolution, such as 3 m or less. IKONOS-2 belongs to this category and can be found in several studies of benthic habitat mapping [41], [289], [333], [456], but mostly before 2015, the year it has

ceased operating. GaoFen-2 satellite, launched in 2014, has the same spatial and spectral resolution as IKONOS-2, but is not as widely used [432], perhaps because of its age: it was launched in 2014, when some sensors already had a better resolution. GaoFen have different satellites (from GaoFen-1 to GaoFen-14) that have the same or a lower resolution than GaoFen-2.

With a similar sensor and a slightly better resolution than IKONOS-2, the Quickbird-2 satellite provides images for several studies of reef mapping [18], [189], [190], [283], [351], [371], [372]. Please note that the Quickbird-2 program was stopped in 2015. Similar features are proposed by the Pleiades-1 satellites, from the Optical and Radar Federated Earth Observation program, also present in the literature [34], [93]. An even higher accuracy can be found with GeoEye-1 satellite, providing images at a resolution of less than 1m, making it particularly useful to study coral reefs [180].

The most common and most precise satellite images come from WorldView satellites. For instance, WorldView-2 (WV-2), launched in 2009, has been widely used for benthic habitat mapping and coastline extraction [79], [91], [212], [261], [279], [292], [387], [413], [432], [436], [447]. Despite the high-resolution images provided by WV-2, the highest quality images available at the current time come from WorldView-3 (WV-3), launched in 2014 [90], [303], [316], [385]. WV-3 has a total of 16 spectral bands and is thus able to compete with hyperspectral sensors with more than a hundred bands (such as Hyperion). Moreover, its spatial resolution is the highest available among current satellites, and is even similar to local measurement techniques such as Unmanned Airborne Vehicles (UAV) [89]. Among all the spectral bands offered by the WV-3 sensors, the coastal blue band (400–450 nm) is especially useful for bathymetry, as this wavelength penetrates water more easily and may help to discriminate seagrass patterns [226]. Although the raw SWIR resolution is lower than the one achieved in visible and near-infrared bands, it can be further processed to generate high-resolution SWIR images [231]. In addition, the WV-3 panchromatic resolution is 0.3 m, which almost reaches the typical size of coral reef elements (0.25 m), thus making it also useful for reef monitoring [71].

To further evaluate the importance of each satellite in the global literature (not only on coral studies) and to detect trends in their use, we searched in Scopus and analyzed the number of articles in which they appear between 2010 and 2020. Several trends can be seen. First, among low-resolution satellites, it appears that while the usage of Landsat remains stable over the year, the usage of Sentinel has exploded (by a multiplication factor of 20 between the period 2014–2014 and 2018–2020). Regarding high-resolution satellites, we detect trends in their usage: in the period 2010–2014, Quickbird and IKONOS satellites were predominant, but their usage decreased by more than 85% during the years 2018–2020. On the other hand, the number of papers published using WorldView and PlanetScope has been increasing: respectively from 108 and 0 in 2010–2014, to 271 and 164 in 2018–2020. The complete numbers for each satellite can be found in Figure 2.A.1.

Figure 2.3 depicts which satellites were employed in the 37 studies using satellites ("satellite only" and "satellite + other" in Figure 2.2). Please note that some studies use data from more than one satellite. From this analysis, WorldView satellites appear to be the most commonly used ones for coral mapping, confirming that high-resolution multispectral satellites are more suitable than low-resolution ones for coral mapping.

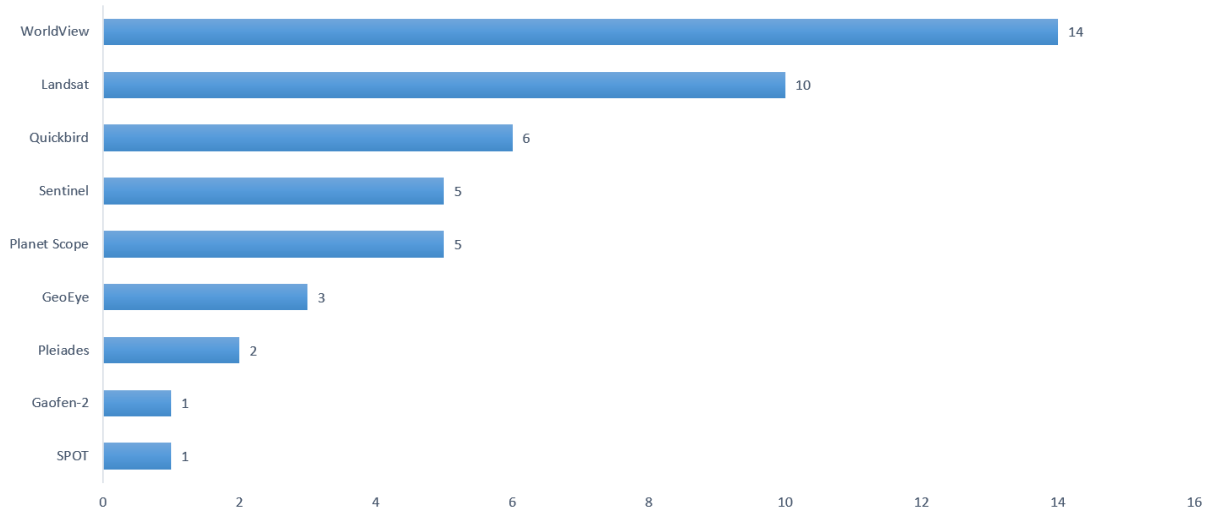


Figure 2.3: Most used satellites in coral reef classification and mapping between 2018 and 2020.

2.3 Image Correction and Preprocessing

Even though satellite imagery is a unique tool for benthic habitat mapping, providing remote images at a relatively low cost over large time and space scales, it suffers from a variety of limitations. Some of these are not exclusively related to satellites but are shared with other remote sensing methods such as UAV. Most of the time, existing image correction methods can overcome these problems. In the same way, preprocessing methods often result in improved accuracy of classification. However, the efficiency of these algorithms is still not perfect and can sometimes induce noise when trying to create coral reef maps. This part will describe the most common processing that can be performed, as well as their limitations.

2.3.1 Clouds and Cloud Shadows

One major problem of remote sensing with satellite imagery is missing data, mainly caused by the presence of clouds and cloud shadows, and their effect on the atmosphere radiance measured on the pixels near clouds (adjacency effect) [118]. For instance, Landsat-7 images have on average a cloud coverage of 35% [211]. This problem is globally present, not only for the ocean-linked subjects but for every study using satellite images, such as land monitoring [56], [379] and forest monitoring [186], [193]. Thus, several algorithms have been developed in the literature to face this issue [187], [195], [208], [262], [366], [408], [453], [461]. One widely used algorithm for cloud and cloud shadow detection is Function of mask, known as Fmask, for images from Landsat and Sentinel-2 satellites [125], [340], [463]. Given a multiband satellite image, this algorithm provides a mask giving a probability for each pixel to be cloud, and performs a segmentation of the image to segregate cloud and cloud shadow from other elements. However, the cloudy parts are just masked, but not replaced.

A common approach to remove cloud and clouds shadows is to create a composite image from multi-temporal images. This involves taking several images at different time periods but close enough to assume that no change has occurred in between, for instance over a few weeks [15]. These images are then combined to take the best cloud-free parts of each image to form one final composite image without clouds nor cloud shadows. This process is widely

used [44], [129], [248], [417] when a sufficient number of images is available.

2.3.2 Water Penetration and Benthic Heterogeneity

The issue of light penetration in water occurs not only with satellite imagery, but with all kinds of remote sensing imagery, including those provided by UAV or boats. The sunlight penetration is strongly limited by the light attenuation in water due to absorption, scattering and conversion to other forms of energy. Most sunlight is therefore unable to penetrate below the 20 m surface layer. Hence, the accuracy of a benthic mapping will decrease when the water depth increases [134]. The light attenuation is wavelength dependent, the stronger attenuation being observed either at short (ultraviolet) or long (infrared) wavelengths while weaker attenuation in the blue-green band allows deeper penetration. Specific spectral bands such as the green one may be viable for benthic habitat mapping and coral changes, such as bleaching [240]. As the penetration through water depends on the wavelength, image preprocessing may be needed to correct this effect. Water column correction methods enable retrieval of the real bottom reflectance from the reflectance captured by the sensor, using either band combination or algebraic computing depending on the method used. Using a water column correction method can improve the mapping accuracy by more than 20% [290], [305]. Several models of water column correction exist, each of them with different performances [464], the best known one being Lyzenga's [258]. The best model strongly depends on the input data and the desired result; see Zoffoli et al. 2014 [464] for a detailed overview of the water column correction methods.

When it is known that the water depth of the study field is homogeneous, it is possible to classify the benthic habitat without applying any correction [12]. However, even in a shallow environment that would be weakly impacted by the light penetration issue (i.e., typically less than 2m deep), a phenomenon called spectral confusion can occur if the depth is not homogeneous [179]. At different depths, the response of two different-color elements can be similar on a wide part of the light spectrum. Hence, with an unknown depth variation, the spectral responses of elements such as dead corals, seagrasses, bleached corals and live corals can be mixed up and their separability significantly affected, making it harder to map correctly [437]. Nevertheless, this depth heterogeneity problem can be overcome: when mixing satellite images with *in situ* measurements (such as single-beam echo sounder), it is possible to have an accurate benthic mapping of reefs with complex structures in shallow waters [385]. However, the advantage of not needing ground-truth data (information collected on the ground) when working with satellite imagery is lost with this solution.

2.3.3 Light Scattering

When remotely observing a surface such as water, especially with satellite imagery, its reflectance may be influenced by the atmosphere. Two phenomena modify the reflectance measured by the sensor. First, the Rayleigh's scattering causes smaller wavelengths (e.g., blue 400 nm) to be more scattered than larger ones (e.g., red 800 nm). Secondly, small particles present in the air cause so-called aerosol scattering, also altering the radiance perceived by the satellites [126], [173]. Hence, the reflectance perceived by the satellite's sensors is composed of the true reflectance to which are added both Rayleigh- and aerosol-related scattered components [156], [158].

It is possible to apply algorithms to correct the effects due to Earth's atmosphere [155], [157], [282], making some assumptions such as the horizontal homogeneity of the atmosphere, or the flatness of the ocean. However, these atmospheric corrections do not always result in a significant increase in the classification accuracy when using multispectral images [249], and they are not as frequent as water column corrections, which is why we consider them as optional.

2.3.4 Masking

Masking consists of removing geographic areas that are not useful or usable: clouds, cloud shadows, land, boats, wave breaks, and so on. Masking can improve the performance of some algorithms such as crop classification [121] or Sea Surface Temperature (SST) retrievals [224].

Even though highly accurate algorithms exist to detect most clouds, as discussed previously, some papers employ manual masking for higher accuracy [153]. It is also possible to mask deep water, in which coral reef mapping is difficult to achieve. Deep water to be masked can be defined by a criterion such as a reflectance threshold over the blue band (450–510 nm) [66].

2.3.5 Sunglint Removal

When working with water surfaces, such as an ocean or lagoon, sunglint poses a high risk of altering the quality of the image, not only for satellite imagery but for every remote sensing system. Sunglint happens when the sunlight is reflected on the water surface with an angle similar to the one the image is being taken with, often because of waves. Thus, higher solar angles induce more sunglint; on the other hand, they are also correlated with a better quality for bathymetry mapping based on physical analysis methods [153]. Although this reflectance can be easily avoided when taking field images from an airborne vehicle (by controlling the time of the day and the direction), it is harder to avoid with satellite imagery. It thus must be removed from the image for better accuracy of benthic habitat mapping. This can be achieved, for instance, by a simple linear regression [176]. Some other models can also efficiently tackle this issue [112], [113], [215], [272]. According to Muslim et al. 2019 [291], the most efficient sunglint removal procedure when mapping coral reef from UAV is the one described in Lyzenga et al. 2006 [259]. As the procedures compared in the paper depend on multispectral UAV data, we can imagine that the result may be true for satellite data as well.

2.3.6 Geometric Correction

Geometric correction consists of georeferencing the satellite image by matching it to the coordinates of the elements on the ground. It allows, for instance, removal of spatial distortion from an image or drawing a parallel between two different sources of data, such as several satellite images, satellite images with other images (e.g., aerial), or images mixed with bathymetry inputs (sonar, LiDAR). This step is especially important in the case of satellite imagery, which is subject to a large number of variations such as angle, radiometry, resolution or acquisition mode [140].

Geometric corrections are needed to be able to use ground-truth control points. These data can take several forms, for instance divers' underwater videos or acoustic measurements from a boat. Even though control points are not used in every study, they are frequent because they enable a high-quality error assessment and/or a more accurate training set. However, control points are not that easy to acquire because they require a field survey, which is not always possible and may be expensive for some remote sites. Thus, control points are not always used.

2.3.7 Radiometric Correction

When working with multi-temporal images of the same place, the series of images is likely to be heterogeneous because of some noise for instance induced by sensors, illumination, solar angle or atmospheric effects. Radiometric correction enables normalization of several images to make them consistent and comparable. Radiometric corrections significantly improve accuracy in change detection and classification algorithms [80], [313], [419], [450]. Identically to geometric corrections that are only required when working with ground-truth control points, radiometric corrections are only needed when working with several images of the same place. Radiometric corrections are also useful to provide factors needed in the equations of some atmospheric correction algorithms [459].

2.3.8 Contextual Editing

Contextual editing is a postprocessing of the image, subsequent to the classification step that takes into account the surrounding pattern of an element [163], [441]. Indeed, some classes cannot be surrounded by another given class, and if it is found to be the case then the classifier has probably made a mistake. For instance, an element classified as "land" that is surrounded by water elements is more likely to be a class such as "algae".

The use of contextual editing can greatly enhance the performance of a classifier, be it for land area [404] or for coral reefs [45], [290]. However, surprisingly, it appears that this method has not been widely employed in the published literature, especially with benthic habitat related topics. To the best of our knowledge, even though we found some papers using contextual editing for bathymetry studies, it has not been applied to coral reef mapping in the past 10 years.

2.4 From Images to Coral Maps

Satellite imagery represents a powerful tool to assess coral maps, should we be able to tackle the problems that come with it. Manual mapping of coral reefs from a given image is a long and arduous work and synthetic expert mapping over large spatial area and/or long time periods is definitely out of reach, especially when the area to be mapped has a size of several km^2. Coral habitats are at the moment unequally studied, with some sites that are almost not analyzed at all by scientists: for instance, studies on cold-water corals mostly focus on North-East Atlantic [246]. The development of automated processing algorithms is a necessary step to target a worldwide and long-term monitoring of corals from satellite images. The mapping of coral reefs from remote sensing usually follows the flow chart given in Andréfouët 2008 [17] consisting of several steps of image corrections, as seen previously,

followed by image classification. For instance, with one exception, all the studies published since 2018 that deal with mapping coral reefs from satellite images perform at least three out of the four preprocessing steps given in [17]. The following subsections provide a comparison of the accuracies given by different statistical and machine-learning methods.

2.4.1 Pixel-Based and Object-Based

Before comparing the machine-learning methods, a difference must be drawn between two main ways to classify a map: pixel-based and object-based. The first consists of taking each pixel separately and assigning it a class (e.g., coral, sand, seagrass, etc.) without taking into account neighboring pixels. The second consists of taking an object (i.e., a whole group of pixels) and giving it a class depending on the interaction of the elements inside of it.

The object-based image analysis method performs well for high-resolution images, due to a high heterogeneity of pixels which is not suited for pixel-based approaches [43]. This implies that object-based methods should be used in the study of reef changes working with high-resolution multispectral satellite images instead of low-resolution hyperspectral satellite images. Indeed, the object-based method has an accuracy 15% to 20% higher than the pixel-based one in the case of reef change detection [8], [66], [460] and benthic habitats mapping [20], [406].

The relative superiority of the object-based approach has also been shown when applied to land classification [128], [341], such as bamboo mapping [139] or tree classification [97], [201]. Nonetheless, even if the object-based methods are generally more accurate, they remain harder to set up because they need to perform a segmentation step (to create the objects) before the classification.

2.4.2 Maximum Likelihood

Maximum likelihood (MLH) classifiers are particularly efficient when the seabed does not have a too complex architecture [438]. With good image condition, i.e., clear shallow water (<7 m) and almost no cloud cover, a MLH classifier can discriminate *Acropora* spp. corals from other classes (sand, seagrass, mixed coral species) with an accuracy of 90% [65]. Moreover, a MLH classifier works well under two conditions: when the spectral responses of the habitats are different enough to be discriminated, and when the area analyzed is in shallow waters (< 5 m) [456]. It is however very likely that these results can be applied to other machine-learning methods.

Nevertheless, when compared to other classification methods such as Support Vector Machine (SVM) or Neural Networks (NN), MLH classifiers appear to be less efficient, be it for land classification [3], [219], [220], [249], [411] or for coastline extraction [181], [279]. A comparison of some algorithms applied to crop classification also confirms that SVM and NN perform better, with an accuracy of more than 92% [229].

2.4.3 Support Vector Machine

Across several studies, SVM appears to be the method with the best accuracy [3], [219], [431], especially with edge pixels, i.e., pixels which border two different classes [181]. In the studies published between 2018 and 2020, SVM classifiers had on average an accuracy of 70% for coral mapping, but can achieve up to 93% classification accuracy among 9 different

classes of benthic habitat [162] when coupling high-resolution satellite images from WV-3 with drone images [162].

2.4.4 Random Forest

Random Forest (RF) methods also are very efficient in remote sensing classification problems [42]. They perform well to classify and map seagrass meadows [166] or land-use [46], the most often forests [263], [297], [392], although performance for land ecosystems may not be directly compared to that obtained for marine habitat mapping. For shallow water benthic mapping, a RF classifier can still outperform a SVM classifier in benthic habitat mapping [438], or at least have an identical overall accuracy but with a better spatial distribution. Globally, RF classifiers can map benthic habitat with an overall accuracy ranging from 60% to 85% [4], [26], [235], [328], [440], [457], depending on the study site, the satellite imagery involved, and the preprocessing steps applied to the images.

2.4.5 Neural Networks

NN are commonly used to classify coral species with underwater images, but to date have rarely been used to map coral reefs from satellite images alone [7], [239], [432]. However, NN often appear in papers where satellite images are mixed with other sources such as aerial photographs, bathymetry data, or underwater images [82], [278]. NN can be useful to perform a segmentation to extract features before performing the classification with a more common machine-learning method such as SVM [432] or K-nearest neighbors, with more than 80% mapping accuracy [239].

2.4.6 Unsupervised Methods

Unsupervised machine-learning methods are less frequent but still appear in some studies. The most present methods are based on K-means and ISODATA [202], [357], [406] with an accuracy ranging from 50% to 80%, reaching 92% when discriminating between 3 benthic classes [29]. The latter is an improvement of the former, where the user does not have to specify the number of clusters as an input. In the first place, the algorithm clusters the data, and then assign each cluster a class.

2.4.7 Synthesis

Given that the results on which is the best classifier can vary from a paper to another, we decided to gather in Figure 2.4 the coral mapping studies since 2018 using satellite imagery only. Please note that we excluded the methods that appeared in less than 3 papers, leading to an analysis of a subset of 20 study of the 25 papers depicted in Figure 2.2. We regrouped the methods K-means and ISODATA under the same label "K-means +" because these two methods are based on the same clustering process. Despite our comprehensive search of the literature, we acknowledge the possibility that some studies may have been overlooked. All the papers used here can be found in Table 2.B.1.

From the previous section and Figure 2.B.1, we recommend that the most accurate methods are RF and SVM. However, this recommendation has to be carefully evaluated because all the studies compared in this paper are based on different methods (how the

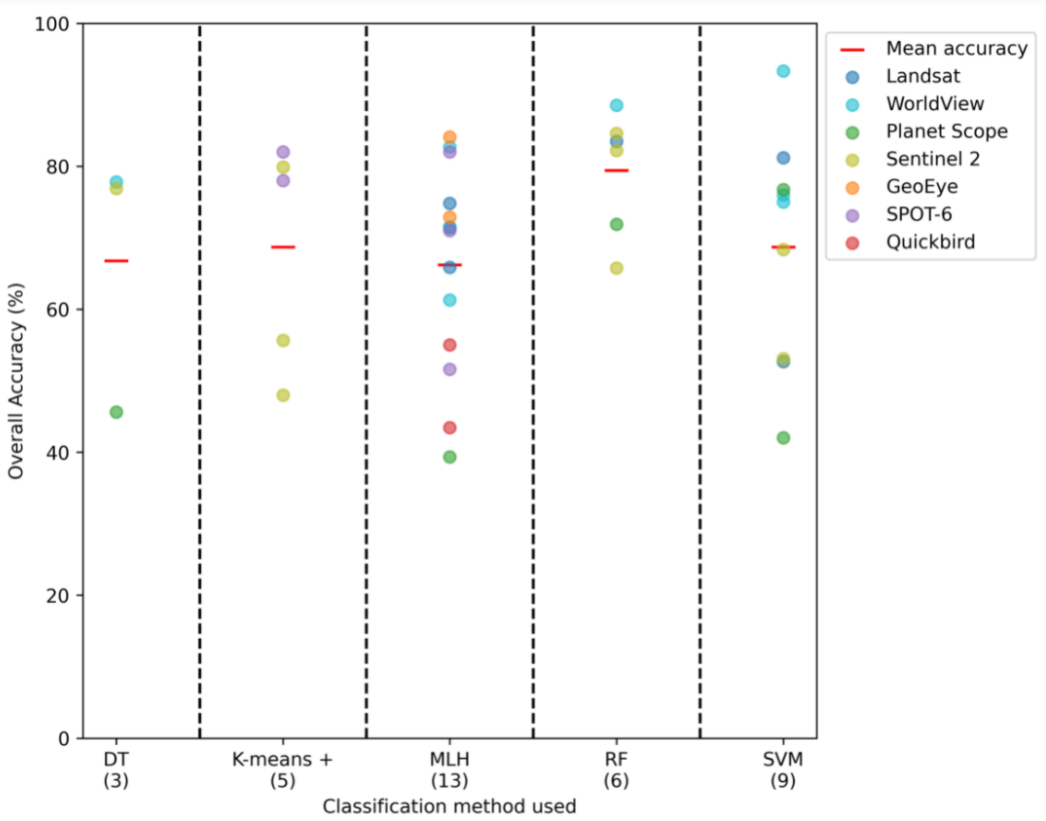


Figure 2.4: Accuracy of 20 studies from 2018 to 2020 depending on the method and satellite used. One point is a one study, its X-axis value correspond to the method used and its color correspond to the satellite used. One paper can create several points if it used different methods or different satellites. The red line is the mean of each method. The method "K-means +" regroups the methods K-means and ISODATA. "RF" is Random Forest, "SVM" is Support Vector Machine, "MLH" is Maximum Likelihood and "DT" is Decision Tree. The number of studies using each method appears in parentheses.

performance of the model is evaluated and which preprocessing are performed on the images) and data sets (location of the study site and satellite images used), which may have a strong influence on the obtained results.

2.5 Improving Accuracy of Coral Maps

Although we have focused so far on satellite images-derived maps, there are many other ways to locate coral reefs without directly mapping them. This section will describe how to study reefs without necessarily mapping them, and the technologies that allow improvements in the precision of reef mapping.

2.5.1 Indirect Sensing

It is possible to acquire information on reefs and their localization without directly mapping them. Indirect sensing refers to these methods, studying reefs by analyzing their surrounding factors.

For instance, measuring Sea Surface Temperatures (SST) have helped to draw conclusions that corals have already started adapting to the rise of ocean temperature [253]. Similarly,

as anomalies in SST are an important factor in coral outplant survival [120], an algorithm forecasting SST can predict which heat stress may cause a coral bleaching event [250]. Furthermore, it is possible to use deep neural networks to predict SST even more accurately [364]. However, even though the measured SST and the real temperature experienced by reefs can be similar [148], it is not always the case depending on the sensors used and other measurements such as wind, waves and seasons [426]. To try to overcome this issue and obtain finer predictions of the severity of bleaching events, it is possible to combine water temperature with other factors such as the light stress factor [388], known to be a cause of bleaching [237].

Backscatter and absorption measurements, as well as chlorophyll-a levels, can also be analyzed to detect reef changes [218], [390]. The chlorophyll-a levels and total suspended matter can be highly accurately retrieved with some algorithms based on satellite images [24], [98], [274], [381]. Similarly, computation on bottom reflectance can detect coral bleaching [448].

We could imagine that these indirect measurements, performed with satellite imagery and providing useful data about coral health, could be incorporated as an additional input to some classifiers to improve their accuracy. This is something we have not been able to find in current literature and that we suggest trying.

2.5.2 Additional Inputs to Coral Mapping

First, to enhance the classification accuracy, it appears evident that a higher satellite image resolution implies a higher accuracy for a same algorithm [275]. Notwithstanding this, we will describe here the different means to enhance the mapping with a given satellite resolution.

To be able to effectively detect environmental changes, several factors are important [288], among which the quality of the satellite images [99] and the quantity of data over time [223]. Indeed, it is essential to have a temporal resolution of a few days or even less, to be able to select the best images, without cloud nor sunglint [327]. A solution can thus be to couple images from a high-resolution satellite with a high-frequency satellite, for instance WV-3 and RapidEye [162].

To be able to discriminate some coral reefs with a special topography, satellite imagery may not be enough. Adding bathymetry data, for instance acquired with LiDAR, can improve the accuracy of the results [41], [66], [94], [267], [276]. It is possible to estimate bathymetry and water depth, with one of a numbered methods that currently exist [92], [185], [244], [293], and to include this as an additional input to a coral reef mapping algorithm [95]. This method is found in Collin et al. 2021 [90], where it improves the accuracy by up to 3%, allowing more than 98% overall accuracy with high-resolution WV-3 images.

Underwater images can also be used jointly with satellite images. They can be obtained from underwater photos taken by divers [76], [351], [376], as well as underwater videos taken from a boat [284].

To conclude, we recommend mixing several input data to improve accuracy: photo transects, underwater camera videos, bathymetry, salinity or temperature measurements [32], [138], [170], [338], [352].

2.5.3 Citizen Science

Crowd sourcing can help classify images or provide large sets of data [81], [122], [238], [252], [269], [362], [403], in remote sensing of coral reefs as well as in other fields. However, the citizen scientists can be wrong or provide different classification [61], [84], [277] and thus still some modifications are often needed to learn from citizens' responses [207], [363]. The Neural Multimodal Observation and Training Network (NeMO-Net), a NASA project, is a good example of how citizen science can be used to generate highly accurate 3D maps and provide a global reef assessment, based on an interactive classification game [83], [239], [423]. This type of data can especially be helpful to feed a neural network, knowing that ground-truth knowledge and expert classification are hard to acquire.

2.6 Conclusions and Recommendations

Through all the papers studying coral reefs between 2018 and 2020 and mapping them from satellite imagery, the best results are obtained with RF and SVM methods, even though the achieved overall accuracy almost never reaches 90%, and is often below 80%. The arrival of very high-resolution satellites dramatically increases this to more than 98% [90]. To map coral reefs with a higher accuracy, we recommend using satellite images with additional inputs when it is possible.

When performing coral mapping from satellite images, it is very common to apply a wide range of preprocessing. Out of the four preprocessing methods proposed in Andréfouët 2008 [17], we suggest applying a water column correction (see [464] for the best method), and a sunglint correction (we recommend [259]). Geometric correction is only needed when working with ground-truth points, and radiometric correction when working with multi-temporal images. Interestingly, some postprocessing methods such as contextual editing appear to be less well used and could improve accuracy [45], [290].

Presently, several projects exist to study and map coral reefs at a worldwide scale, using an array of resources, from satellite imagery to bathymetry data or underwater photographs: the Millennium Coral Reef Mapping Project [13], the Allen Coral Atlas [32] or the Khaled bin Sultan Living Ocean Foundation [57]. These maps are proven useful to the scientific community for coral reef and biodiversity monitoring and modeling, as well as inventories or socio-economic studies [14].

However, when examining the maps created by all these projects, we can see that many sites are yet to be studied. Furthermore, some reef systems have been mapped at a given time but would need to be analyzed more frequently, to be able to detect changes and obtain a better understanding of the current situation. Hence, even if the work achieved to date by the scientific community is huge, a lot still needs to be done. Great promise lies in upcoming very high-resolution satellites coupled with the cutting-edge technology of machine-learning algorithms.

Appendix 2.A

Figure 2.A.1 depicts the number of articles in which each satellite appears in Scopus, for three different periods: 2010–2014, 2015–2017, 2018–2020.

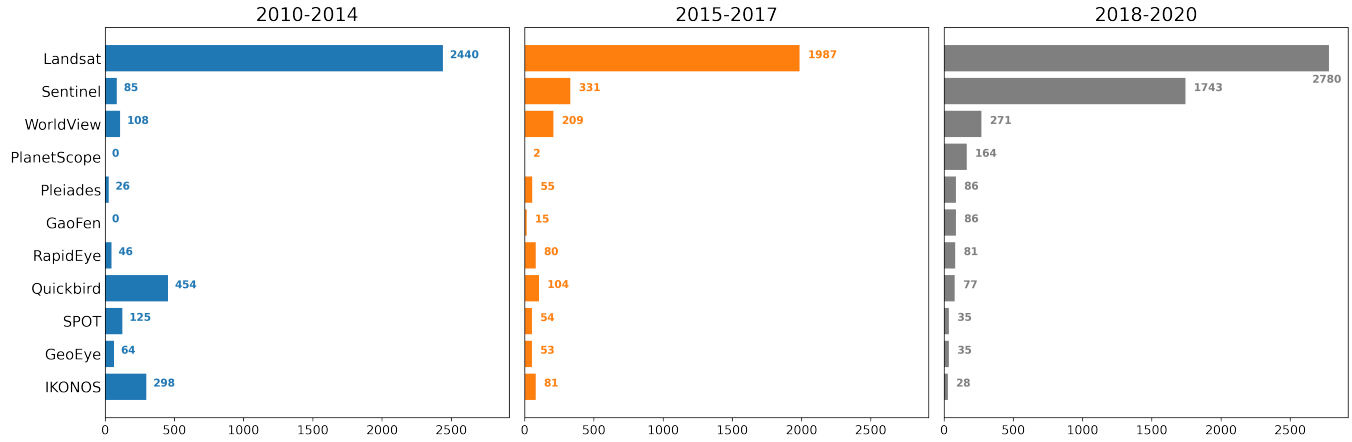


Figure 2.A.1: Number of articles in which each satellite appears in the Scopus database, depending on the years.

Appendix 2.B

Table 2.B.1 summarizes the 20 studies that have been used to build Figure 2.4.

Table 2.B.1: Studies from 2018 to 2020 used to compare the accuracies of different methods.

Reference	Satellite Used	Method Used	Nb. of Classes
Ahmed et al. 2020 [4]	Landsat	RF, SVM	4
Anggoro et al. 2018 [20]	WV-2	SVM	9
Aulia et al. 2020 [30]	Landsat	MLH	6
Fahlevi et al. 2018 [115]	Landsat	MLH	4
Gapper et al. 2019 [133]	Landsat	SVM	2
Hossain et al. 2019 [189]	Quickbird	MLH	4
Hossain et al. 2020 [190]	Quickbird	MLH	4
Immordino et al. 2019 [202]	Sentinel-2	ISODATA	10 and 12
Lazuardi et al. 2021 [235]	Sentinel-2	RF, SVM	4
McIntyre et al. 2018 [275]	GeoEye-1 and WV-2	MLH	3
Naidu et al. 2018 [292]	WV-2	MLH	7
Poursanidis et al. 2020 [328]	Sentinel-2	DT, RF, SVM	4
Rudiastuti et al. 2021 [357]	Sentinel-2	ISODATA, K-Means	4
Shapiro et al. 2020 [378]	Sentinel-2	RF	4
Siregar et al. 2020 [387]	WV-2 and SPOT-6	MLH	8
Sutrisno et al. 2021 [406]	SPOT-6	K-Means, ISODATA, MLH	4
Wicaksono & Lazuardi 2018 [439]	Planet Scope	DT, MLH, SVM	5
Wicaksono et al. 2019 [438]	WV-2	DT, RF, SVM	4 and 14
Xu et al. 2019 [447]	WV-2	SVM, MLH	5
Zhafarina & Wicaksono 2019 [457]	Planet Scope	RF, SVM	3

Conclusion

In this chapter, a thorough examination of recent papers investigating the mapping of coral reefs using multispectral satellite images has been conducted. The analysis reveals that the SVM and RF methods often yield the most promising outcomes, thus guiding the thesis towards this specific direction.

Moreover, preprocessing methods have emerged as a common tool that has demonstrated its ability to enhance the results. However, it is important to note that certain preprocessing techniques, such as incorporating bathymetry knowledge, may necessitate additional data. Notably, the inclusion of bathymetry data has proven to be valuable in improving the accuracy of mapping outcomes. Nevertheless, due to several reasons that will be detailed in the last chapter of this thesis, the images used to develop our final workflow did not undergo preprocessing steps, with the exception of corrections inherent in the acquired images.

Furthermore, it has been observed that higher resolution satellite images generally yield superior performances in coral reef mapping. However, the acquisition of such high-resolution imagery incurs greater expenses.

Additionally, the integration of supplementary data sources, including underwater images or temperature measurements, has the potential to further enhance mapping accuracy. However, it is essential to highlight that this thesis specifically focuses on utilizing solely satellite image inputs. By adopting this approach, the aim is to provide a user-friendly tool that eliminates the need for fieldwork.

Finally, despite the existence of certain mapping projects, it is important to note that these maps either lack global coverage or exhibit accuracy issues in specific regions. Furthermore, these maps are generated based on a fixed timeframe and remain static thereafter, limiting their practicality. In contrast, our objective is to develop a framework that can generate maps for any given input image, thus ensuring broader applicability and flexibility.

CHAPTER 3

SMOTE for compositional data

The work in this chapter is a paper published in PLoS ONE: T. Nguyen, K. Mengersen, S. Meulé, D. Sous and B. Liquet. SMOTE-CD: SMOTE for compositional data. *PLoS ONE*, 18(6):e0287705, 2023 [299].

The work in this chapter also lead to the creation of the Python package *smote-cd*, released on PyPi: https://pypi.org/project/smote-cd.

Synopsis

During the mapping process of an image, when trying to perform a classification on the polygons created by a segmentation method, the labels end up being compositional. Besides, imbalance in the labels may arise as the number of pixels belonging to each class may differ. Hence, we are facing the issue of imbalance problem within compositional data, which no method currently exists to deal with.

Based on the Synthetic Minority Oversampling TEchnique (SMOTE), this chapter provides a new method to oversample compositional data, that we called SMOTE for Compositional Data (SMOTE-CD). The efficiency of this technique is assessed by comparing the performances of three regressors (Gradient Boosting tree, Neural Networks, Dirichlet regressor) on a synthetic dataset and on Maupiti dataset.

What is important to note here is that the Maupiti dataset used in this chapter is a simplified version of the final dataset, as it gathered the originally fifteen classes into four classes, forgetting some deep areas of the lagoon. This has been made in the seek of simplicity, faster results, and most importantly to obtain presentable results for the Dirichlet regressor which produced extremely low performances on the full dataset.

Abstract

Compositional data are a special kind of data, represented as a proportion carrying relative information. Although this type of data is widely spread, no solution exists to deal with the cases where the classes are not well balanced. After describing compositional data imbalance, this paper proposes an adaptation of the original Synthetic Minority Oversampling TEchnique (SMOTE) to deal with compositional data imbalance. The new approach, called SMOTE for Compositional Data (SMOTE-CD), generates synthetic examples by computing a linear combination of selected existing data points, using compositional data operations. The performance of the SMOTE-CD is tested with three different regressors (Gradient Boosting tree, Neural Networks, Dirichlet regressor) applied to two real datasets and to synthetic generated data, and the performance is evaluated using accuracy, cross-entropy, F1-score, R2 score and RMSE. The results show improvements across all metrics, but the impact of oversampling on performance varies depending on the model and the data. In some cases, oversampling may lead to a decrease in performance for the majority class. However, for the real data, the best performance across all models is achieved when oversampling is used. Notably, the F1-score is consistently increased with oversampling. Unlike the original technique, the performance is not improved when combining oversampling of the minority classes and undersampling of the majority class. The Python package *smote-cd* implements the method and is available online.

3.1 Introduction

3.1.1 Context

Over the past few years, data imbalance problems have been widely studied in classification tasks [168]. An imbalance distribution over the classes will often cause the models to prioritize their performance on the majority classes, at the expense of the minority ones. Different methods exist to deal with imbalanced datasets [209]: algorithm-level methods, where the algorithm reduces the bias by inducing a weight on the classes; data-level methods, where the data are modified to reach a more balanced state; and hybrid methods, combining both algorithm-level methods and data-level methods. Among data-level methods, Synthetic Minority Oversampling TEchnique (SMOTE) [75], with all its variations [119], is one of the most popular for classification problems. The SMOTE algorithm generates synthetic data points for a particular class by combining the features of two existing points belonging to the same class through linear interpolation.

Most algorithms designed to tackle class imbalance problems, such as SMOTE, are often limited to the classification tasks; for instance [53], [78], [222], [320]. However, even though regression problems are also very common in real-life problems, only a few resampling strategies exist for regression tasks [319], [414].

In this paper, we address the special issue of dealing with an imbalanced dataset in regression problems in the case where the labels are compositional. Compositional data are data carrying relative information [5], presented as proportions or percentages, making them different from other types of data. Compositional data are encountered in various fields, including biology [380], [418], [446], chemistry [2], [124], ecology [205], [429], geology [58], [87], and social sciences [116], [434], [435], among others. However, the class imbalance

problem in compositional data regression remains a major challenge in the development of effective models. Existing adaptations of SMOTE and other oversampling techniques have focused on addressing imbalanced datasets in single-label regression [69], [194], [286], [415], multi-label classification [74], [104], or when the features are compositional data [159]. However, to the best of our knowledge, no oversampling technique exists for addressing the issue of class imbalance in multi-label regression problems with compositional labels. Therefore, we propose a new oversampling technique called SMOTE for Compositional Data (SMOTE-CD), specifically designed to address this particular situation.

Here, we will measure class imbalance by summing the values of the labels (probability values) for each class on the whole dataset, and summarizing it as a percentage. In that sense, in a perfectly balanced dataset, the percentage of the sum of each class would be $1/K$, with K being the number of classes.

The proposed method is evaluated using five different performance metrics, including accuracy, cross-entropy, F1-score, R2 score, and RMSE, to three different models (Gradient Boosting tree, Neural Networks, Dirichlet regressor) on both simulated and real datasets. Since no other oversampling algorithm currently exists for compositional data, the evaluation of SMOTE-CD is limited to comparing its performance against the case where no oversampling technique is applied. The results show that the performance of the models is overall greater when applying SMOTE-CD, thus demonstrating the effectiveness of the proposed method. This is an important contribution to the field, as it provides a solution for dealing with compositional data imbalance, which has not been addressed before. The use of five different evaluation metrics, as well as the application of three different models to both simulated and real datasets, further strengthens the reliability and generalizability of the proposed method.

The entire paper is arranged as follows. The paper's first section introduces the proposed method and the motivation example. Section 2 presents the compositional data and the SMOTE-CD algorithm. Section 3 presents the metrics, the simulation study and its results. Sections 4 and 5 present the result on the real datasets. Section 6 presents the discussion and conclusion.

3.1.2 Motivation example: Maupiti island

Description of Maupiti island

The overall purpose of our research project is to develop an automated mapping tool able to provide a classification map from a given satellite image, with a particular focus on a coral reef-lagoon system. The test field site is the Maupiti island, the westernmost Leeward island of the Society archipelago, French Polynesia. The site has a size of approximately 8km by 8km. Maupiti data, that we use here, is just an example, but compositional data can be found, for instance, in health or chemistry fields.

An expert-based mapping of Maupiti island was used as a training dataset to develop the model. The satellite image used is a 4-band image (blue, green, red, near infrared) captured on June, 14 2021 by the Pleiades satellite. The expert-based mapping of the image relies on the combination of several field observation campaigns [394] and direct examination of the satellite image. The present analysis focuses on the shallow regions of the lagoon, displaying more interpretable imaging. In the selected areas, four seabed type classes were established (Fig 3.1a):

- **Class 1: Coral**, marked by a overwhelming dominance of coral reef cover.
- **Class 2: Sand**, describing areas covered by detritic sand.
- **Class 3: Shorereef**, gathering shore reef and transitional shore reef.
- **Class 4: Mixed**, representing area covered by a combination of sand and coral.

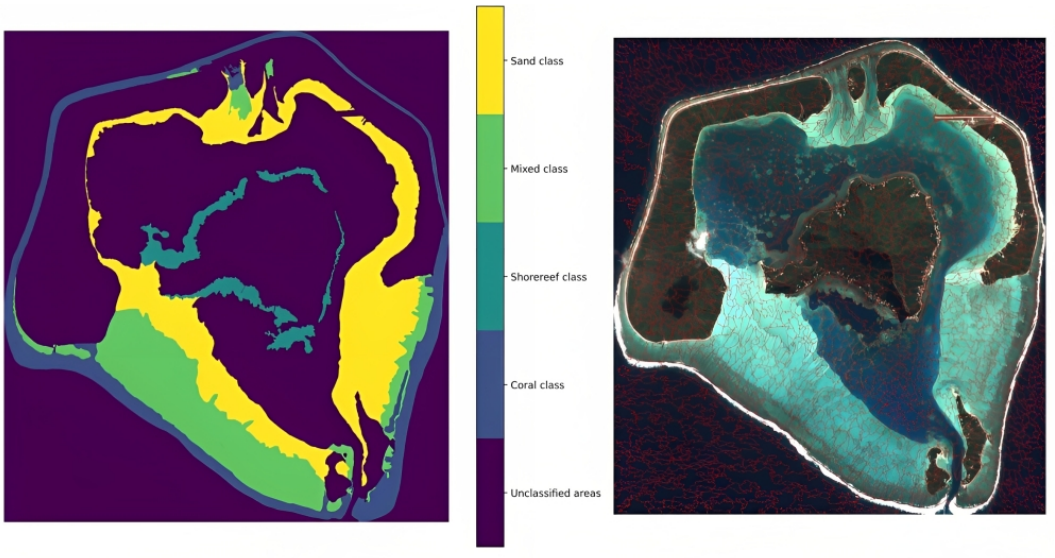
Table 3.1: Percentage of the number of pixels of each class on Maupiti data, based on expert mapping.

Class	Class 1	Class 2	Class 3	Class 4
Percentage	0.117	0.040	0.482	0.361

Automatic mapping

To perform the automatic mapping, the image was first segmented using Felzenszwalb's method [117], which gives Fig 3.1. For each segment, two different operations were applied:

- The four statistical moments (mean, variance, skewness, kurtosis) were computed on each band; these 16 values will be the features of the dataset.
- The percentage of pixels belonging to each class were computed, according to the expert-based classification; this results in a vector that sums up to 1 that will be the labels of the dataset.



(a) (b)

Figure 3.1: (a) Expert-based mapped image of Maupiti island and (b) Pleiades image of Maupiti island segmented with Felzenszwalb's method.

To be able to map the satellite image, the idea was to train a regressor to retrieve, for each segment, the percentage of pixels belonging to each class (i.e., a vector of probabilities).

As shown in Table 3.1, the data are not balanced: one of the class represents 49.5% of the dataset, while another one represents only 3.6%. To overcome this issue, we developed an oversampling technique in order to improve the performance of the regression model on this special kind of data.

3.2 Materials and method

3.2.1 Compositional data

Mathematically, we define a D-part compositional dataset as a vector $x = (x_1, x_2, \ldots, x_D) \in \mathbb{R}^D$ such that,

$$\begin{cases} x_i \geq 0, \quad \forall\, i \in \{1, 2, \ldots, D\}, \\ \sum_{i=1}^{D} x_i = 1. \end{cases}$$

A simplex S^D is defined as the ensemble of all the D-part compositional data, i.e.

$$S^D = \left\{ x = (x_1, x_2, ..., x_D) \mid \forall i \in \{1, 2, \ldots, D\}, x_i \geq 0; \sum_{i=1}^{D} x_i = 1 \right\}.$$

The operations performed in S^D must be adapted to follow the properties of the simplex [5]. For instance, before performing the Euclidian operations, it is possible to first apply the centred log-ratio transform $clr(\cdot)$ to the data,

$$\begin{aligned} clr\colon S^D &\to \mathbb{R}^D \\ (x_1, \ldots, x_D) &\mapsto \left(\log\left(\frac{x_1}{g(x)}\right), \ldots, \log\left(\frac{x_D}{g(x)}\right) \right). \end{aligned}$$

where the function $g(\cdot)$ is the geometric mean $g(x) = \left(\prod_{i=1}^{D} x_i\right)^{\frac{1}{D}}$. The $clr(\cdot)$ function is only defined for vectors where none of the value is equal to 0. Several methods exist to overcome this issue [367], but in practice we just replace the 0 by a tiny value such as 10^{-20}. The definition of the $clr(\cdot)$ function involves the existence of the inverse function $clr^{-1}(\cdot)$, that turns to be the softmax function, defined for $z = (z_1, \ldots, z_D) \in \mathbb{R}^D$ as

$$\text{softmax}(z) = \frac{1}{\sum_{i=1}^{D} \exp(z_i)} \cdot (\exp(z_1), \ldots, \exp(z_D)).$$

It is also possible to directly define operators on S^D. Let C be the closure operator,

$$\forall k \in \mathbb{N}, C(x_1, \ldots, x_k) = (x_1, \ldots, x_k)/(x_1 + \cdots + x_k).$$

For two D-part compositions $x, y \in S^D$, the perturbation $x \oplus y$ is defined by

$$x \oplus y = C(x_1 y_1, \ldots, x_D y_D), \tag{3.1}$$

and, given $\alpha \in \mathbb{R}$, the power transformed composition $\alpha \otimes x$ is

$$\alpha \otimes x = C(x_1^\alpha, \ldots, x_D^\alpha). \tag{3.2}$$

3.2.2 SMOTE for compositional data

In this section, we denote by n the number of samples in the dataset, p the number of features and K the number of classes. The matrix $X \in \mathbb{R}^{n \times p}$ contains the n observations of the p features and $Y \in \mathbb{R}^{n \times K}$ contains their labels. For any $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, K\}$, we denote by $y_{i,j}$ the value of Y at row i and column j, and $y_{i,\cdot} = (y_{i,1}, \dots, y_{i,K})$ the probability vector label of row i. Similarly, for $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, p\}$, $x_{i,j}$ is the value of X at row i and column j, and $x_{i,\cdot} = (x_{i,1}, \dots, x_{i,p})$. In order to simplify the notation, we define

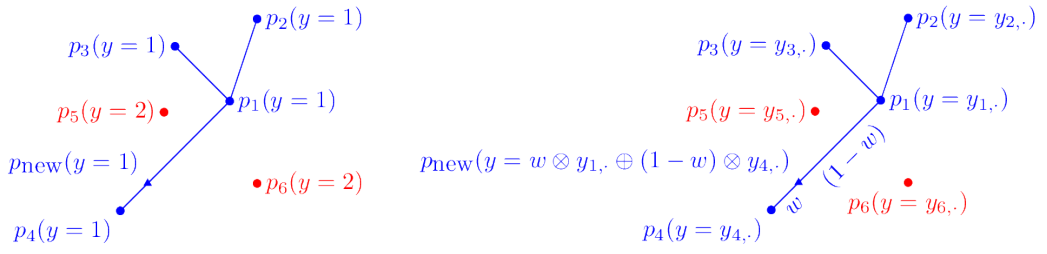
$$\operatorname{argmax}(y_{i,\cdot}) = \operatorname*{argmax}_{j \in \{1,\dots,K\}} (y_{i,j}),$$

which represents the majority class of a given label $y_{i,\cdot} \in [0,1]^K$. We also define the sum vector $S \in \mathbb{R}^K$ as the sum of the values for each class,

$$S = \left(\sum_{i=1}^{n} y_{i,1}, \sum_{i=1}^{n} y_{i,2}, \dots, \sum_{i=1}^{n} y_{i,K} \right). \tag{3.3}$$

The majority class of the dataset is thus defined as $\operatorname{argmax}(S)$, and the minority class as $\operatorname{argmin}(S)$.

Before introducing the SMOTE-CD algorithm, let's first summarize the idea behind the original SMOTE algorithm. As shown in Fig. 3.2(a), the SMOTE algorithm creates a new point that belongs to class 1 (represented by blue points). To achieve this, the algorithm first selects a point at random (in this case, p_1) and identifies its nearest neighbors (p_2, p_3, p_4). Note that only neighbors with the same label as p_1 (i.e., class 1) are considered, while points labeled as class 2 (represented by red points) are ignored. The algorithm then chooses one of these neighbors (p_4) and creates a new point along the line that connects p_1 and p_4. The features of the new point are determined through a linear combination of the features of p_1 and p_4, and its label is assigned as 1. Algorithm 1 describes the SMOTE algorithm.



(a) Original SMOTE.

(b) SMOTE for Compositional Data.

Figure 3.2: Difference between the original SMOTE algorithm and SMOTE-CD. The blue points are the points to oversample. (a) The points to oversample belong to the same class (here, class 1). (b) The points to oversample are the ones that have the same class as their majority class in their compositional vector label.

The SMOTE-CD algorithm keeps the main ideas from the original SMOTE: 1) select a point from the class to be oversampled, 2) select one of its k-Nearest Neighbors ($k \in \mathbb{N}$ specified by the user) and 3) create a synthetic point in-between those two points. Because of the label that is compositional, these three steps have to be adapted:

1. Select a point r_1 whose majority class is m, where m is the minority class of the dataset.

Algorithm 1 Original SMOTE [75]

Require: $X \in \mathbb{R}^{n \times p}$ the features.
Require: $Y \in \{1, \dots, J\}^n$ the class label outputs.
Require: $k \in \mathbb{N}$ the number of neighbors to select for the k-Nearest Neighbors.
Ensure: Generated data $X_{\text{new}} \in \mathbb{R}^{q \times p}$ and $Y_{\text{new}} \in \{1, \dots, J\}^q$ with q the number of points created.

```
1: Denote by S_j the number of points labeled as class j.
2: M ← the majority class of dataset.
3: Initialize X_new and Y_new as empty matrices.
4: for every class m that needs to be oversampled do
5:    while S_m < S_M do
6:       Compute D = {i | y_i = m}, the set of points labeled as class m.
7:       Randomly choose r_1 ∈ D and find the indices of its k nearest neighbors.
8:       Randomly choose an index r_2 among these neighbors.
9:       x^new ← w × x_{r_1,·} + (1 − w) × x_{r_2,·} with w ∈ [0, 1] randomly drawn.
10:      y^new ← m.
11:      S_m ← S_m + 1.
12:      Append x^new to X_new, append y^new to Y_new.
13:   end while
14: end for
15: return X_new, Y_new
```

2. Compute the k-Nearest Neighbors of r_1 among the points that also have m as their majority class. Then select a point r_2 in one of these k neighbors.

3. Randomly draw $w \in [0, 1]$. The features of the point to be created is a linear combination of the two points selected before, with w being the weight of r_2 and $(1 - w)$ the weight of r_1. Similarly, the labels of the point to be created is a linear combination, but using the operators from Eq. (3.1) and (3.2).

Fig. 3.2(b) depicts an example of how SMOTE-CD creates a new point. As we are dealing with compositional data label, every point p_i has a vector label y_i. All the blue points are the points having the class m as the majority class of their label y_i, where m is the minority class of the dataset. The algorithm computes the 3 nearest neighbors of p_1 only considering the blue points, and then a point is created on the line between p_1 and p_4. The label of the new point is a linear combination of the labels y_1 and y_4 using the operations defined on the simplex (Eq. (3.1) and (3.2)).

Algorithm 2 describes the SMOTE-CD algorithm, using the same notation.

The step that creates the label of the new point (line 12) uses the definitions of Eq (3.1) and (3.2). Nevertheless, it is also possible to create the label by using the Euclidian operations on the logratio transformed labels, and to apply the inverse transformation afterwards: $clr^{-1}\left(w \times clr(y_{r_1,\cdot}) + (1 - w) \times clr(y_{r_2,\cdot})\right)$. Although the label could be created by directly performing Euclidian operations on the compositional label, however this would be mathematically irrelevant because it would not respect the rules of compositional data analysis [6].

The proof of convergence holds in the fact that, at each iteration, the increase of the major class of S is smaller that the increase of its minor one, causing the sum of the minor class to converge to the sum of the major one. In other words, we have to be assured that, at each iteration, $y_m^{\text{new}} > y_M^{\text{new}}$, with m (resp. M) the minority (resp. majority) class of the dataset.

Algorithm 2 SMOTE for compositional data

Require: $X \in \mathbb{R}^{n \times p}$ the features.
Require: $Y \in \mathbb{R}^{n \times K}$ the labels (compositional data).
Require: $k \in \mathbb{N}$ the number of neighbors to select for the k-Nearest Neighbors.
Ensure: Generated data $X_{\text{new}} \in R^{q \times p}$ and $Y_{\text{new}} \in R^{q \times K}$ with q the number of points created.

Compute the label sum vector $S \in \mathbb{R}^D$ as defined in Eq (3.3).
M $\leftarrow$ argmax(S), the majority class of dataset (hence S_M is the sum of the majority class).
Initialize X_{new} and Y_{new} as empty matrices.
while min$(S) < S_M$ **do**
    $m \leftarrow$ argmin(S), the minority class of dataset.
    Compute $\mathcal{D} = \{i \mid \text{argmax}(y_{i,\cdot}) = m\}$, the set of points whose majority class is m.
    Randomly choose an index $r_1 \in \mathcal{D}$.
    Find the indices of the k nearest neighbors of r_1 in $\mathcal{D}$, using the Euclidian distance on X.
    Randomly choose an index r_2 among these indexes.
    Uniformly draw a number $w \in [0, 1]$.
    $x^{\text{new}} \leftarrow w \times x_{r_1,\cdot} + (1 - w) \times x_{r_2,\cdot}$.
    $y^{\text{new}} \leftarrow w \otimes y_{r_1,\cdot} \oplus (1 - w) \otimes y_{r_2,\cdot}$.
    $S \leftarrow S + y^{\text{new}}$.
    Append x^{new} to X_{new}, append y^{new} to Y_{new}.
end while
return X_{new}, Y_{new}

This is straightforward by using Eq. (3.2) and (3.1) to write y^{new} as:

$$y^{\text{new}} = C\left(\frac{y_{r_1,1}^w}{\sum_j y_{r_1,j}^w} \cdot \frac{y_{r_2,1}^{1-w}}{\sum_j y_{r_2,j}^{1-w}}, \ldots, \frac{y_{r_1,J}^w}{\sum_j y_{r_1,j}^w} \cdot \frac{y_{r_2,J}^{1-w}}{\sum_j y_{r_2,j}^{1-w}}\right),$$

and then noticing that the two indices r_1 and r_2 used for generating a new point are chosen in $\mathcal{D} = \{i \mid \text{argmax}(y_{i,\cdot}) = m\}$:

$$\begin{aligned} r_1, r_2 \in \mathcal{D} &\Rightarrow \begin{cases} y_{r_1,m} > y_{r_1,M} \\ y_{r_2,m} > y_{r_2,M} \end{cases} \\ &\Rightarrow \begin{cases} y_{r_1,m}^w > y_{r_1,M}^w \\ y_{r_2,m}^{1-w} > y_{r_2,M}^{1-w} \end{cases} \\ &\Rightarrow y_{r_1,m}^w y_{r_2,m}^{1-w} > y_{r_1,M}^w y_{r_2,M}^{1-w} \\ &\Rightarrow y_m^{\text{new}} > y_M^{\text{new}}. \end{aligned}$$

3.3 Simulation study

3.3.1 Data simulation

The simulated data are generated by using a multinomial logistic regression. The main idea is to create a probability distribution from a multinomial logistic regression, and then use a Dirichlet distribution with those probabilities to generate the actual label of the new point.

The notation is the same as in the previous section : the number of features (resp. classes) is p (resp. K), and the number of samples is n. The user has to specify a matrix $B \in [0, 1]^{(p+1) \times K}$ which corresponds to the regression coefficients, where $B_{i,k}$ is associated

with the ith feature and the kth class. For instance, for a class k, the regression coefficients will be $(B_{0,k}, B_{1,k}, \ldots, B_{p,k})$. Note that $B_{0,k}$ is the intercept, hence explaining the $(p+1) \times K$ dimension of B.

For a given point $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$, we define $x' = (1, x_1, \ldots, x_p) \in \mathbb{R}^{p+1}$ and a vector α as:

$$\begin{aligned}\alpha &= \text{softmax}(B_{0,1} + B_{1,1}x_1 + \cdots + B_{p,1}x_p, \ldots, B_{0,K} + B_{1,K}x_1 + \cdots + B_{p,K}x_p) \\ &= \text{softmax}(x' \cdot B_{\cdot,1}, \ldots, x' \cdot B_{\cdot,K}).\end{aligned}$$

We are then able to randomly draw a label for x with a Dirichlet distribution with parameter α. Algorithm 3 generates a random dataset using this method.

Algorithm 3 Function to generate a synthetic dataset with compositional labels

Require: $K \in \mathbb{N}$ the number of classes.
Require: $p \in \mathbb{N}$ the number of features.
Require: $n \in \mathbb{N}$ the number of samples.
Require: $B \in [0,1]^{(p+1)\times K}$ the regression coefficients, where $B_{m,k}$ is associated with the mth feature and the kth class.
Ensure: Generated data $X \in R^{n\times p}$ and $Y \in R^{n\times K}$
1: Create a random matrix of points $X \in R^{n\times p}$ such that for all i, j, $x_{i,j}$ is a random number uniformly drawn in a chosen interval (for instance [-10, 10])
2: Initialize Y as an empty matrix of size $(n \times K)$.
3: **for** every row x in X (and its associated row index i) **do**
4: Compute $\alpha = \text{softmax}(x' \cdot B_{\cdot,1}, \ldots, x' \cdot B_{\cdot,K})$ where $x' = (1, x_1, x_2, \ldots, x_p)$
5: Randomly draw a vector from a Dirichlet distribution with parameter α and attribute it to $y_{i,\cdot}$, the ith row of Y.
6: **end for**
7: **return** X, Y

To better understand how the regression coefficients B can change the configuration of the data, we give an example of simulated data with 2 features and 2 labels. Two different values $B^{(a)}$ and $B^{(b)}$ are tested :

$$B^{(a)} = \begin{bmatrix} 0.4 & 0.4 \\ 0.2 & 0.4 \\ 0.5 & 0.3 \end{bmatrix}, \qquad B^{(b)} = \begin{bmatrix} 0.1 & 0.9 \\ 0.0 & 0.5 \\ 0.8 & 0.1 \end{bmatrix}.$$

Each column of a matrix B represents the coefficients for one class. There are 3 lines here because there are 2 features and the first value corresponds to the intercept of the regression. In $B^{(a)}$, the coefficients of each class are purposely close to each other, while they are easily separable in $B^{(b)}$. Fig 3.3 shows the value of the labels when generating the same 400 points with each matrix, using the function *generate_dataset* of our *smote-cd* Python package, with *random_state*=2. The points created with $B^{(b)}$ have a clearer border between the points fully belonging in one class or the other. As there are only two classes and their sum is 1, it is only necessary to represent the value of one of them with the gradient of color.

3.3.2 Performance measures

The value of row i column j of Y is still denoted by $y_{i,j}$, and is the probability that the ith sample belongs to class j. Let $\hat{y}_{i,j}$ be the estimate of this probability by a model.

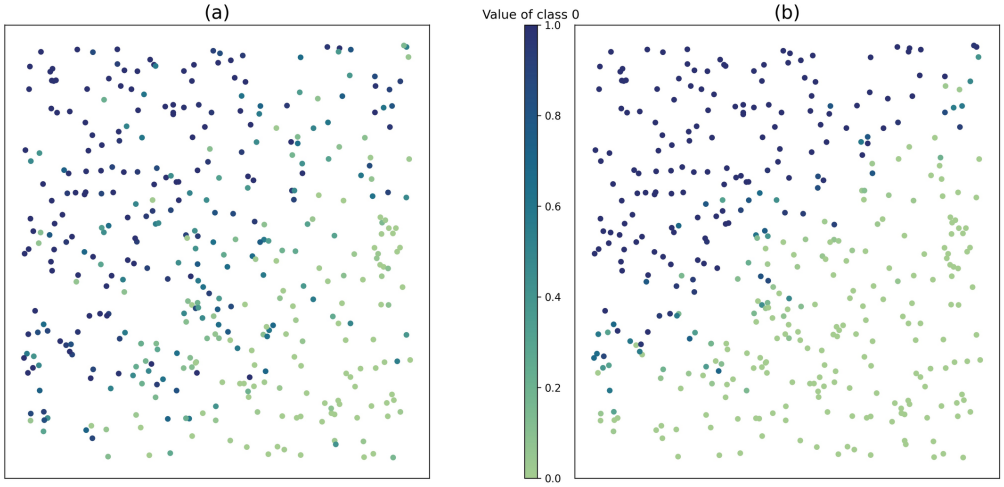


Figure 3.3: Simulation of 400 points using $B^{(a)}$ (a) and $B^{(b)}$ (b).

Different metrics can be used to measure the performance of the model. A popular metric is the cross-entropy:

$$CrossEntropy = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K} y_{i,j}\log(\hat{y}_{i,j}+\varepsilon). \tag{3.4}$$

The ε is added here to overcome the case where $\hat{y}_{i,j} = 0$. We chose $\varepsilon = 10^{-20}$. As the cross-entropy is a loss function, the smaller it is, the better the model performs. The cross-entropy loss may not always be suitable for our model because it treats each sample as equally important, without taking into account the imbalance of the test set. For instance, consider a model predicting three different classes (1, 2 and 3), and imagine that this model performs quite well on class 1 but poorly on classes 2 and 3. If the test set is imbalanced and has a large proportion of class 1 samples, the cross-entropy loss of this model will be low even though it performs poorly overall. The coefficient of determination R^2 allows assessment of the performance of a model on each of the K classes. For a class j, the coefficient of determination is given by

$$R_j^2 = 1 - \frac{\sum_{i=1}^{n}(y_{i,j}-\hat{y}_{i,j})^2}{\sum_{i=1}^{n}(y_{i,j}-\bar{y}_j)^2},$$

where $\bar{y}_j$ is the mean of the values of the jth class. The final R^2 will be equal to the average of the R_j^2 for each class j.

In addition, we also use the Root Mean Squared Error (RMSE) to measure the accuracy of the models. Since we are dealing with multi-class compositional vectors, we define the RMSE between a true and estimated vector as the average of RMSEs calculated across all their classes. Specifically, this is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\frac{1}{K}\sum_{i=1}^{n}\sum_{j=1}^{K}(y_{i,j}-\hat{y}_{i,j})^2}.$$

Even though we are working on a regression problem, classification metrics can be a good tool to understand the efficiency of the models. To do so, it is easy to transform a compositional label $y_{i,\cdot}$ into a class y_i' by applying the argmax,

$$y_i' = \underset{j}{\text{argmax}}\ y_{i,j}.$$

The usual classification metrics can then be applied to y'. Here, we will use the accuracy (the number of correct points divided by the total number of points) and the F1-score which is computed per class,

$$\text{F1-score } = \frac{TP}{TP + \frac{1}{2}(FN + FP)},$$

where TP are the true positive, FN the false negative and FP the false positive. As with the R^2, the F1-score will be computed for each class and then averaged.

3.3.3 Results

First, to investigate the effect of the oversampling technique, synthetic data were generated with 2 features and 2 classes. To make the dataset imbalanced, 90% of the points that had class 0 as a majority class were deleted. We obtain a dataset in which 93% of the points have class 1 as their majority class (Fig 3.4(a)), which is then oversampled by selecting a number of nearest neighbors $k = 10$. Fig 3.4(b) displays the balanced dataset after applying SMOTE-CD, where the original points are displayed as circles and the synthetic created points are displayed as crosses. As in Fig 3.3, the gradient of color represents the value of one of the two classes.

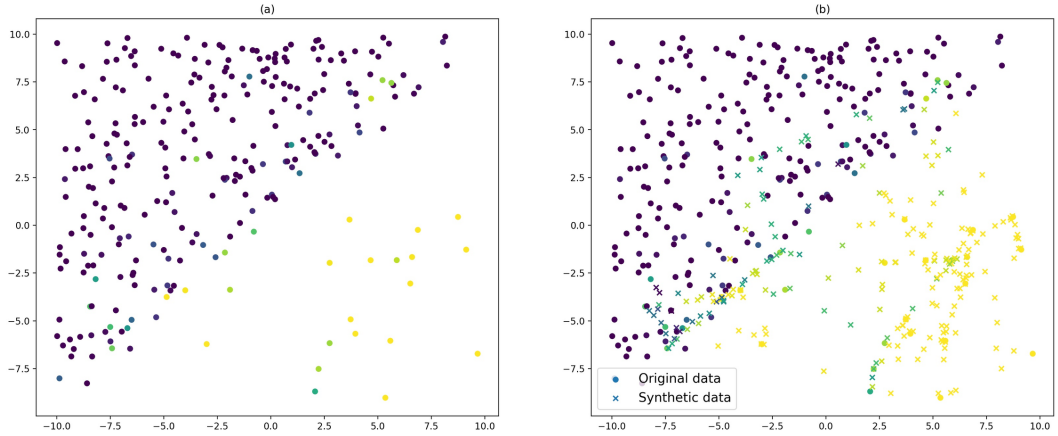


Figure 3.4: An example of SMOTE-CD. (a) The original imbalanced dataset, (b) the output balanced dataset with the created points displayed as a cross.

To evaluate the performance of SMOTE-CD, a 5-fold cross validation was used for three models: Gradient Boosting tree (GB), Neural Network (NN) with one hidden layer, and Dirichlet regression model [265]. The first and second models are chosen because Random Forest and NN are known to be the most efficient to map coral reefs from multispectral satellites [241], [298] and because NN are used in literature for the task of predicting compositional labels [192], [260], and the third is chosen because it is used to generate the simulated data. For each model, the performance is compared between the raw and oversampled data. For the models on which it is possible (GB and NN), hyperparameter tuning was been performed for each data (raw or oversampled). The hyperparameters are detailed in Table S1 and Table S2.

The simulated data were generated with the same shape as the Maupiti data. We selected a matrix B such that the imbalance of the classes was similar to the one of the real data (see Table 3.1). Then, 550 points were created with 16 features and 4 classes to train the models. Testing was performed with 11000 points (20 times the training set size). This operation

was repeated 100 times with the same B. The results and metrics (accuracy, cross-entropy, average F1, RMSE and R^2) are presented in Table 3.2.

Table 3.2: Comparison of simulated raw data (4 classes) and oversampled data, repeated 100 times. Displayed results are mean (s.d.).

	Accuracy	Cross-entropy	F1-score	RMSE	R^2
GB (raw)	0.692 (0.018)	5.272 (1.539)	0.532 (0.045)	0.363 (0.011)	0.137 (0.067)
GB (logratio)	0.724 (0.017)	2.508 (0.553)	0.658 (0.027)	0.341 (0.011)	0.198 (0.074)
GB (compositional)	0.683 (0.016)	3.657 (1.055)	0.604 (0.038)	0.359 (0.011)	0.139 (0.085)
NN (raw)	0.772 (0.026)	3.340 (1.370)	0.611 (0.057)	0.315 (0.020)	0.298 (0.103)
NN (logratio)	0.784 (0.023)	1.700 (0.380)	0.729 (0.033)	0.304 (0.018)	0.301 (0.108)
NN (compositional)	0.750 (0.054)	3.483 (1.367)	0.690 (0.063)	0.332 (0.040)	0.198 (0.207)
Dirichlet (raw)	0.789 (0.016)	0.685 (0.017)	0.605 (0.039)	0.287 (0.004)	0.416 (0.022)
Dirichlet (logratio)	0.875 (0.010)	0.754 (0.017)	0.824 (0.019)	0.303 (0.004)	0.380 (0.022)
Dirichlet (compositional)	0.874 (0.011)	0.755 (0.017)	0.824 (0.019)	0.303 (0.004)	0.379 (0.022)

For both the Gradient Boosting and Neural Network models, the oversampling with logratio distance significantly improves all metrics except for R^2 on the Neural Network ($p < 0.0006$). With the compositional distance on the Neural Network, only the F1-score significantly increases ($p \ll 10^{-10}$), while accuracy, RMSE, and R^2 decrease. The GB model shows significant improvement for cross-entropy, F1-score, and RMSE ($p < 0.008$), but a decrease in accuracy. The Dirichlet model with oversampling significantly increases accuracy and F1-score ($p \ll 10^{-10}$) but decreases cross-entropy, RMSE, and R^2.

In order to understand the effects of the imbalance of the dataset on the performance of the oversampling method, three metrics (accuracy, F1 and R^2) were evaluated with different imbalance ratios. First, a matrix B was created to generate a balanced dataset with 16 features and 4 classes. Then, the ratio of class 0 was increased by incrementing the value of $B_{1,1}$. At each step (for a total of ten steps), the following operation was repeated 100 times : 550 points were created to train the models on the raw or oversampled data, and the models were tested on a set of 11000 points. The result appears in Fig 3.5.

It is apparent that the efficiency of SMOTE-CD depends on the data and the model used. The oversampling technique only improves the R^2 score when the dataset is slightly imbalanced (largest class representing less than 40%), but performs poorly when it is highly imbalanced. On the other hand, the more the dataset is imbalanced, the more the oversampling technique will improve the F1-score. The improvement in accuracy peaks at a certain value of imbalance (when the largest class represents 50% of the dataset), but drops above that threshold.

In order to explain the low R^2 score for the oversampled data, the R^2 per class was calculated for each of the ten steps mentioned above and then averaged. Fig 3.6 displays the result. The average imbalance ratio is 52% for class 0 (and thus approximately 16% for the three other classes).

For the largest class, the R^2 score is decreased from 0.5 to 0.3 by the oversampling technique, which explains why the raw score is higher than the oversampled score in Fig 3.5. However, for the three minority classes, the R^2 is increased by approximately 0.05, which is the initial goal of the method.

Similarly, Fig 3.6 also depicts the F1-score per class, averaged over the seven steps. The difference is that the F1-score of the majority class is not decreased by the oversampling technique, while the score of the minority classes is increased by approximately 0.08.

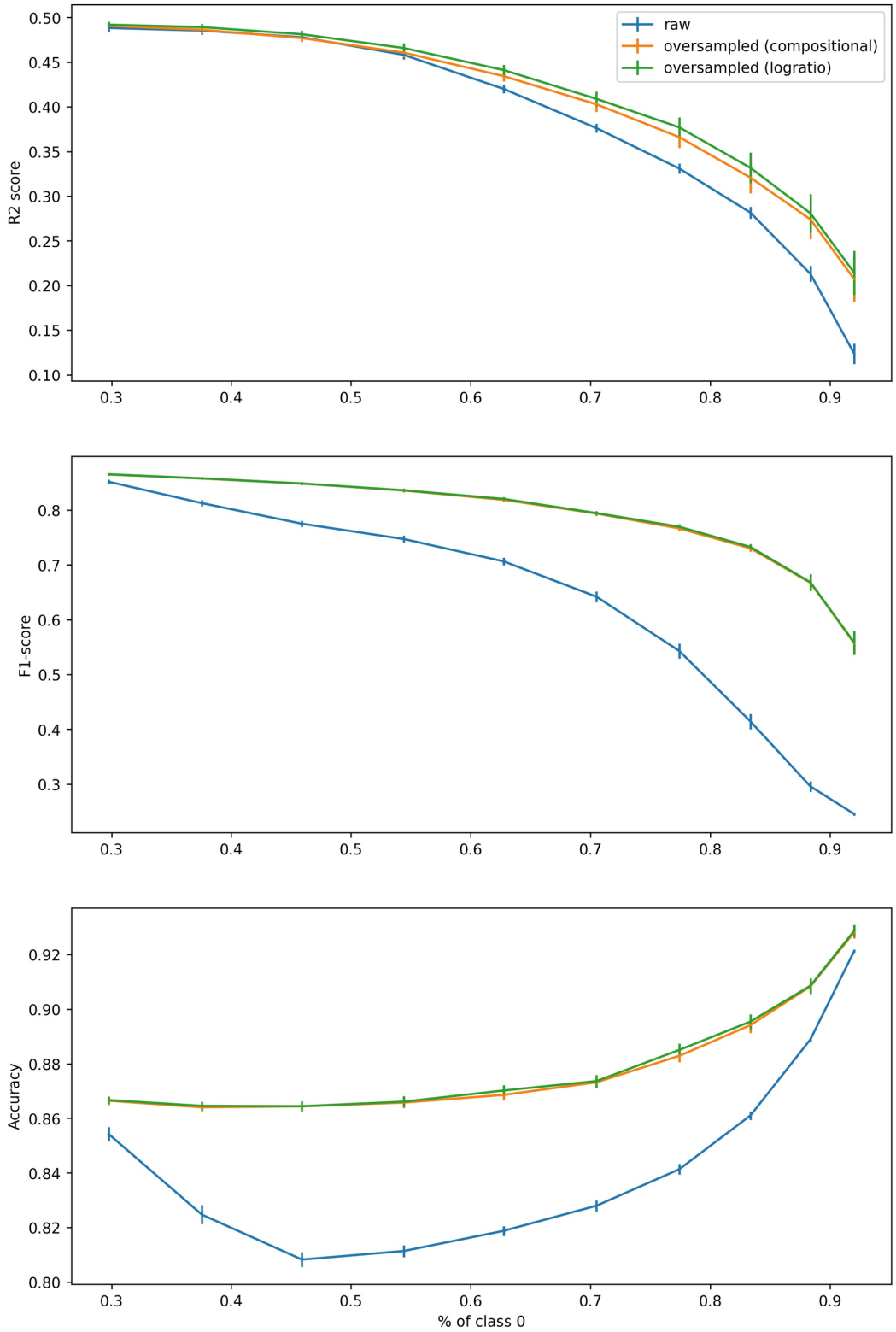


Figure 3.5: Performance of Dirichlet model on raw and oversampled data, depending on the imbalance of the dataset (indicated by % of observations in class 0), based on 16 features and 4 classes.

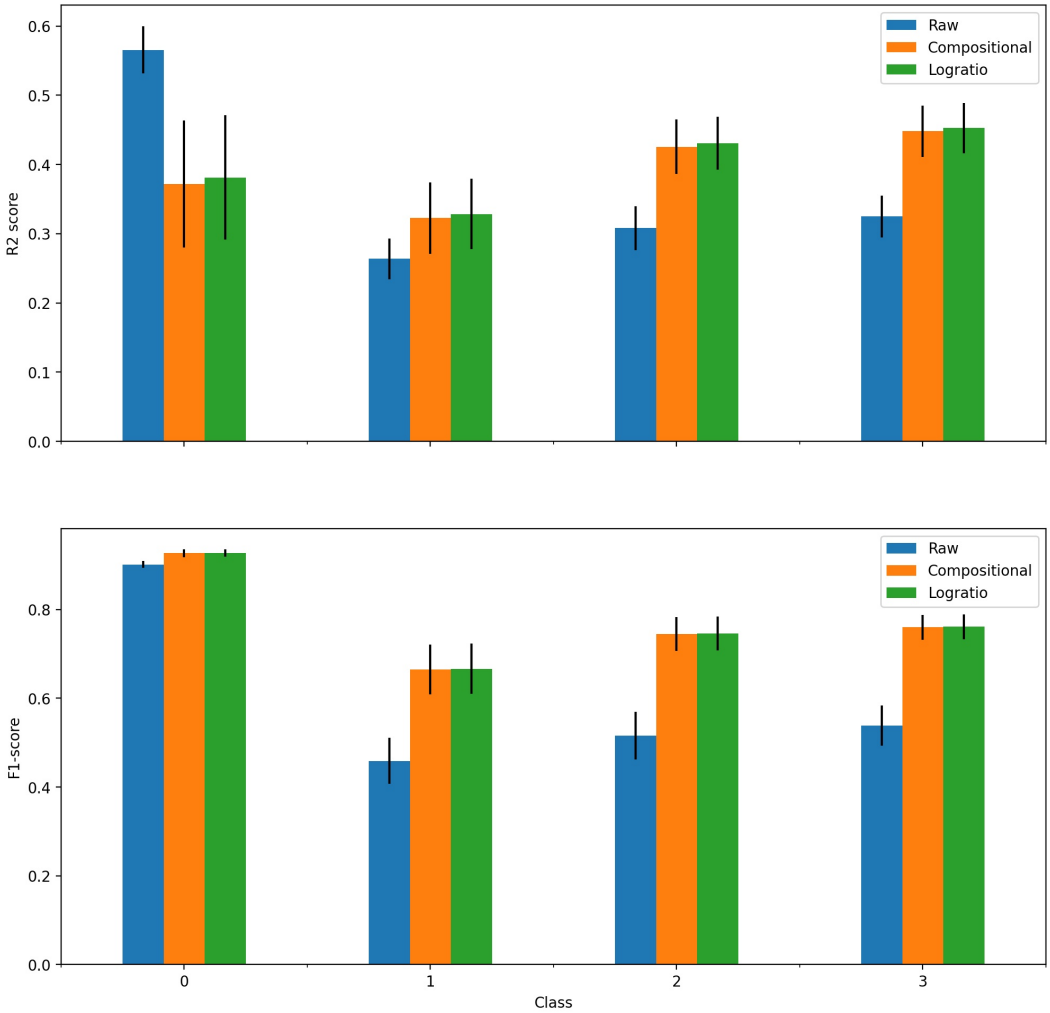


Figure 3.6: Average R^2 and F1-score per class of Dirichlet model on raw and oversampled simulated data. Bars represent the mean score, vertical lines represent the standard deviation.

3.4 Application to Maupiti data

The performance of the three models on the raw dataset was compared with the oversampled dataset (with either the logratio distance used to create the new labels, or the compositional distance). The results are shown in Table 3.3. With the Maupiti dataset, the NN is defined with 2 hidden layers of size 80 and 40, and the relu activation function.

With the GB model, all the metrics are significantly improved ($p < 0.03$) when using the oversampling technique, excepted for the cross-entropy for which the differences are not statistically significant ($p = 0.14$ and $p = 0.38$ respectively for the compositional and the logratio distance). The SMOTE-CD shows less results with the NN and Dirichlet model, where only the difference on the F1 is statistically significant (respectively $p < 0.044$) and $p < 10^{-10}$). This improvement is quite important for the Dirichlet model though, as it represents a difference of almost 0.08.

We analyze the per-class R^2 of the Gradient Boosting tree as it is the best model. Fig 3.7 compares the R^2 between the raw and oversampled data. The oversampling technique decreases the performance of the model for the smallest class (Class 2) for the logratio distance, does not change for the largest class (Class 3) and increases the performance on

Table 3.3: Results comparing raw Maupiti data (4 classes) and oversampled with a 5-fold cross validation. Displayed results are mean (s.d.).

	Accuracy	Cross-entropy	F1-score	RMSE	R^2
GB (raw)	0.857 (0.003)	2.538 (0.196)	0.809 (0.031)	0.229 (0.003)	0.583 (0.018)
GB (logratio)	0.859 (0.003)	2.504 (0.182)	0.822 (0.028)	0.226 (0.003)	0.596 (0.019)
GB (compositional)	0.859 (0.004)	2.486 (0.149)	0.822 (0.028)	0.226 (0.003)	0.596 (0.018)
NN (raw)	0.877 (0.003)	4.048 (0.416)	0.831 (0.008)	0.214 (0.003)	0.624 (0.018)
NN (logratio)	0.877 (0.003)	3.982 (0.456)	0.835 (0.009)	0.214 (0.003)	0.623 (0.017)
NN (compositional)	0.878 (0.003)	3.956 (0.406)	0.834 (0.010)	0.213 (0.003)	0.622 (0.020)
Dirichlet (raw)	0.801 (0.056)	1.676 (0.874)	0.684 (0.127)	0.262 (0.033)	0.420 (0.163)
Dirichlet (logratio)	0.810 (0.049)	1.663 (0.851)	0.762 (0.064)	0.262 (0.036)	0.423 (0.174)
Dirichlet (compositional)	0.810 (0.049)	1.654 (0.839)	0.762 (0.064)	0.262 (0.036)	0.423 (0.174)

the others (Classes 1 and 4).

We conclude that SMOTE-CD does not improve the performance for a class that is too small: in order to perform ideally, it requires enough points to oversample.

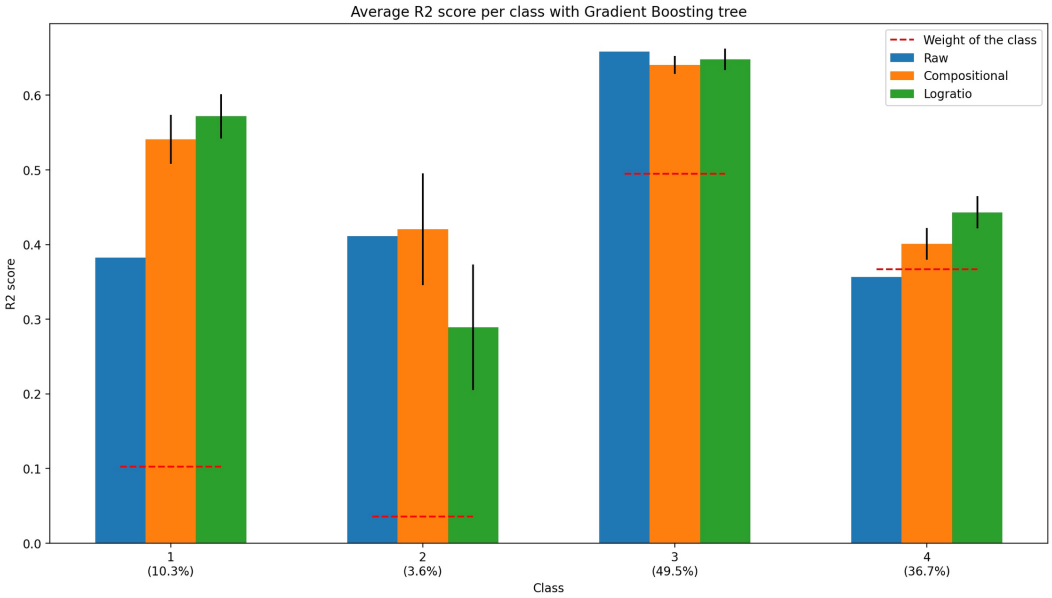


Figure 3.7: Average R^2 score per class of Gradient Boosting tree on raw and oversampled Maupiti data. The red dotted lines represent the weight of each class, and the value below the class is its weight. Bars represent the mean score, vertical lines represent the standard deviation.

3.5 Application to Tecator dataset

To fully evaluate the effectiveness of the SMOTE-CD technique, we applied it to the Tecator meat sample dataset [409], which consists of 240 meat samples. Each sample has absorbance values measured at 100 different wavelengths, as well as corresponding information on the composition of moisture (water), fat, and protein contents. The objective of this analysis is to predict a 3-class compositional data vector from a feature vector of size 100. Because the Dirichlet regression model can be very slow when dealing with a high number of features, we opted to improve its speed by using only the 22 principal components provided in the

dataset instead of the 100 features.

To account for the small size of the dataset, a 10-fold cross validation is applied for each model, iterated over 100 times to vary the folds. The results are displayed in Table 3.4. The neural network is configured with three hidden layers, each having 70 neurons and using the hyperbolic tangent (tanh) activation function, which were selected through hyperparameter tuning.

Table 3.4: Results comparing raw Tecator data (3 classes) and oversampled with a 10-fold cross validation, iterated 100 times. Displayed results are mean (s.d.).

Model	Accuracy	Cross-entropy	F1-score	RMSE	R^2
GB (raw)	0.932 (0.006)	0.860 (0.001)	0.701 (0.042)	0.046 (0.001)	0.717 (0.023)
GB (logratio)	0.957 (0.008)	0.860 (0.001)	0.830 (0.046)	0.044 (0.002)	0.730 (0.027)
GB (compositional)	0.957 (0.008)	0.860 (0.002)	0.834 (0.046)	0.044 (0.002)	0.730 (0.026)
NN (raw)	0.908 (0.000)	0.928 (0.009)	0.512 (0.036)	0.113 (0.005)	-1.230 (0.484)
NN (logratio)	0.904 (0.014)	0.938 (0.010)	0.513 (0.044)	0.122 (0.007)	-1.156 (0.449)
NN (compositional)	0.904 (0.016)	0.937 (0.010)	0.512 (0.044)	0.122 (0.007)	-1.158 (0.466)
Dirichlet (raw)	0.954 (0.007)	0.852 (0.003)	0.846 (0.037)	0.048 (0.003)	0.708 (0.045)
Dirichlet (logratio)	0.940 (0.011)	0.878 (0.006)	0.800 (0.044)	0.072 (0.006)	0.310 (0.224)
Dirichlet (compositional)	0.940 (0.011)	0.877 (0.005)	0.802 (0.037)	0.072 (0.005)	0.323 (0.413)

With the NN, the raw data gives slightly better performances than the oversampled data. However, given the really poor performances of the NN (a negative R^2 and a really high RMSE), we also note that this model was probably not suited for this dataset.

The analysis of the GB and Dirichlet models reveals interesting differences. In both cases, using either the raw or oversampled datasets leads to statistically significant differences ($p < 10^{-4}$). Specifically, for the GB model, using the oversampled data results in better performance, while for the Dirichlet model, oversampling decreases the performance. Notably, among all the models tested, the GB model trained on oversampled data with compositional distance yields the best results. Compared to the Dirichlet model trained on raw data, this approach achieves significantly better accuracy ($p < 0.006$), RMSE ($p \ll 10^{-10}$), and R^2 ($p < 10^{-4}$), with only a slight difference of 1% in cross-entropy and F1-score.

In light of these results, it is apparent that SMOTE-CD can improve the performance for a model that does not perform too poorly (e.g. a R^2 above 0.3). Indeed, if a model has low performance, it is more likely that this is due to poor fit to the data than from the imbalance of the dataset.

3.6 Discussion

The results on the synthetic datasets show that the SMOTE-CD technique can significantly improve the F1-score and accuracy, but it has a mixed effect on other metrics depending on the model and dataset imbalance level. SMOTE-CD improves the overall performance of the model, especially with respect to the accuracy and the F1-score in the cases where the dataset is not too heavily imbalanced. The R^2 score of the majority class remains similar, but the R^2 of a very small class (3% of the dataset) will be decreased. The R^2 of all the other classes is improved, which is the desired goal of the method.

The results on the real datasets show that the SMOTE-CD technique can significantly improve the performance of the Gradient Boosting model for all metrics, while it has a less

pronounced effect on the other models. The per-class analysis of the R^2 score reveals that the SMOTE-CD technique can improve the performance for some classes but not for others, depending on the model and distance metric used.

Further tests are required with other datasets having compositional labels, but these are often hard to find because they are not publicly available. Our oversampling technique could be used with datasets in biology and metabolomics, in poll studies or in soil analysis, but its effectiveness depends on several factors that should be carefully considered.

The original SMOTE paper [75] proposes to undersample the dataset before applying the oversampling technique, which we similarly tested here. The synthetic dataset was first undersampled by randomly withdrawing some points from the majority class, until the total sum of the largest class was equal to the sum of the second largest one. SMOTE-CD was then applied. The results are summarised in Table S3 and compared with those in Table 3.2 when not using undersampling (Table S4). No significant difference can be seen when using undersampling before the oversampling, be it positive or negative. The results are similar when undersampling not only the points having the largest class as their majority class, but the points having one of the n largest classes as their majority class (with $n \in [1, \ldots, 3]$). At this point, we are not able to exclude the utility of the undersampling and suggest it could once more depend on the dataset or on the way the removed points are chosen. For instance, when performing random undersampling, consideration could be given to an Edited Nearest Neighbor approach [443]; see [38].

Work has still to be done regarding the initial selection of the points, because it can influence the performance of the original SMOTE algorithm. For instance, we could imagine attributing a "safe" level to each point by exploring its k nearest neighbors and using it in the creation of a new point [60]. It would also be possible to only oversample the points on the border [172], where the border would here be defined by the points having a given amount of neighbors that have the largest class as their majority class.

3.7 Conclusion

The SMOTE algorithm has been adapted to deal with the special case in which the dataset labels are compositional, which had not been done before. The present study investigates its effectiveness on imbalanced datasets for three different models: Gradient Boosting tree, Neural Networks, and Dirichlet Regression. The evaluation was performed on both synthetic and real datasets, and several metrics, including accuracy, F1-score, RMSE, cross-entropy, and R^2, were used to assess the performance of the models.

The study suggests that the effectiveness of the SMOTE-CD technique depends on several factors, including the model, distance metric, dataset imbalance level, and class distribution. The SMOTE-CD technique can improve the performance of a model that does not perform too poorly, but it may not be effective for a model with very low performance.

An implementation is proposed in the Python package *smote-cd* available on PyPi: https://pypi.org/project/smote-cd. The Jupyter notebooks used to simulate the data and perform the analyses can be found on the GitHub page of the package: https://github.com/teongu/smote_cd.

Appendix 3.A Table S1

Table 3.S1: Hyperparameters of the Gradient Boosting tree. The hyperparameters listed here are those applied to the Gradient Boosting tree of the Python package *scikit-learn*, tuned with the *hyperopt* package. The value of the *random_state* is 2.

	Raw	Oversampled (compositional)	Oversampled (logratio)
ccp_alpha	10	10	0.5
learning_rate	0.01	0.01	0.01
max_depth	5	5	4
max_features	log2	sqrt	sqrt
min_samples_leaf	10	1	10
n_estimators	200	100	200

Appendix 3.B Table S2

Table 3.S2: Hyperparameters of the Neural Networks. The hyperparameters listed here are those applied to the MLPRegressor of the Python package *scikit-learn*, tuned with the *hyperopt* package. The value of the *random_state* is 2.

	Raw	Oversampled (compositional)	Oversampled (logratio)
activation	identity	logistic	identity
alpha	$1e^{-5}$	$1e^{-3}$	$1e^{-3}$
beta_1	0.95	0.95	0.9
hidden_layer_sizes	(40,)	(20,)	(80,)
learning_rate	constant	constant	constant
learning_rate_init	0.001	0.0001	0.0001
max_iter	10000	10000	10000
momentum	0.9	0.8	0.9
solver	sgd	adam	adam

Appendix 3.C Table S3

Table 3.S3: Results comparing simulated raw data (4 classes) and oversampled repeated 100 times, when applying undersampling beforehand.

	R^2	Accuracy	F1-score
GB (raw)	0.141 (0.194)	0.694 (0.061)	0.526 (0.138)
GB (logratio)	0.147 (0.214)	0.707 (0.036)	0.635 (0.087)
GB (compositional)	0.130 (0.255)	0.688 (0.036)	0.600 (0.087)
NN (raw)	0.302 (0.306)	0.773 (0.816)	0.610 (0.173)
NN (logratio)	0.295 (0.311)	0.784 (0.046)	0.727 (0.092)
NN (compositional)	0.212 (0.668)	0.754 (0.158)	0.694 (0.189)
Dirichlet (raw)	0.413 (0.056)	0.781 (0.051)	0.594 (0.102)
Dirichlet (logratio)	0.379 (0.066)	0.874 (0.031)	0.823 (0.056)
Dirichlet (compositional)	0.381 (0.071)	0.874 (0.026)	0.824 (0.056)

Appendix 3.D Table S4

Table 3.S4: Difference when applying undersampling+oversampling, and oversampling only. Results are in bold when the undersampling provides better results.

	R^2	Accuracy	F1-score
GB (logratio)	-0.05	-0.017	-0.022
GB (compositional)	-0.009	**0.006**	-0.003
NN (logratio)	-0.008	0	-0.003
NN (compositional)	**0.019**	**0.006**	**0.005**
Dirichlet (logratio)	-0.002	-0.001	-0.002
Dirichlet (compositional)	**0.002**	0	**0.001**

Conclusion

In this chapter, our primary focus has been on addressing the challenge of data imbalance, with a particular emphasis on compositional data, which play a central role in the context of this thesis.

To tackle this issue, we have introduced a novel oversampling technique explicitly designed for compositional data. Our approach is an adaptation of the Synthetic Minority Oversampling Technique (SMOTE) and has been named SMOTE for Compositional Data (SMOTE-CD).

Through extensive testing involving datasets with three or four distinct classes, we have observed notable enhancements in the performance of the Gradient Boosting model across various evaluation metrics. These evaluations encompassed both synthetic and real-world datasets.

It is important to note that the effectiveness of SMOTE-CD depends on the initial performance of the model. When the model demonstrates low performance at the outset, suggesting a mismatch between the model and the data, the application of SMOTE-CD does not yield significant improvements.

CHAPTER 4

Spatial autoregressive model on a Dirichlet distribution

The work in this chapter is in preparation for a submission to Computational Statistics & Data Analysis: T. Nguyen, S. Moka, K. Mengersen and B. Liquet. Spatial autoregressive model on a Dirichlet distribution.

The work in this chapter has been presented to the MODSIM2023 conference (https://mssanz.org.au/modsim2023/).

Synopsis

The compositional data we work with typically exhibit strong spatial correlation as they represent elements on a geographical map. While methods exist to address spatiality in data, such as Spatial AutoRegressive (SAR) models, there has been a limited focus on applying SAR models to compositional data.

In our approach, we chose to adapt a Dirichlet regression model, primarily due to the suitability of the Dirichlet distribution for handling compositional data. This choice aligns with the fundamental characteristics of compositional data, where each data point represents a composition of parts that sum to a constant, making the Dirichlet distribution a natural choice.

The upcoming chapter introduces a SAR model on a Dirichlet distribution. To highlight the effectiveness of the spatial model, we conduct a comparative analysis against the non spatial model using one synthetic dataset and three real datasets. Through this exploration, we illustrate the superiority of the spatial model in capturing the spatial intricacies inherent in some compositional datasets.

Abstract

Compositional data are widely utilized in various fields, such as ecology, geology, economics, and public health, as they effectively represent proportions or percentages of different components in a whole. Spatial dependencies often exist in compositional data, particularly when the components represent different land uses or ecological variables. Spatial autocorrelation can arise from shared environmental conditions or geographical proximity. Therefore, it is essential to incorporate spatial information into the statistical analysis of compositional data to obtain accurate and reliable results. However, due to correlation between the components of the compositional data, and the constraint of lying on a simplex, traditional statistical methods are not directly applicable to compositional data. To handle compositional data, the Dirichlet distribution is commonly used because its support is a compositional vector. The R package DirichletReg already proposes a regression model for Dirichlet-distributed data, but this model does not consider spatial dependencies, which limits its applicability in spatial problems. In this study, we introduce a spatial autoregressive model for Dirichlet-distributed data that incorporates spatial dependencies between observations. We develop a maximum likelihood estimator on a Dirichlet density function that includes a spatial lag term. To expedite computations, we compute the derivatives and Hessian matrix. We compare this spatial autoregressive model with the same model without spatial lag and test both models on synthetic and two real datasets, using metrics such as R^2, RMSE, cosine similarity, cross-entropy or AIC. By considering the spatial relationships among observations, our model provides more accurate and reliable results for the analysis of compositional data. The model is also compared to a spatial multinomial regression model for compositional data and their respective effectiveness is discussed.

4.1 Introduction

Compositional data are widely used in a range of fields, such as ecology, geology, economics, and public health, due to their ability to represent proportions or percentages of different components in a whole. However, compositional data present a unique challenge to statistical analysis, as they carry relative information and are constrained to lie on a simplex [5]. Thus, traditional statistical methods cannot be applied directly to these types of data.

Mathematically, a D-part compositional dataset is defined as a vector $y = (y_1, y_2, \ldots, y_D) \in \mathbb{R}^D$ such that,

$$\begin{cases} y_i \geq 0, \quad \forall\, i \in \{1, 2, \ldots, D\}, \\ \sum_{i=1}^{D} y_i = 1. \end{cases}$$

A simplex S^D is defined as the ensemble of all the D-part compositional data, i.e.

$$S^D = \left\{ y = (y_1, y_2, ..., y_D) \mid \forall i \in \{1, 2, \ldots, D\}, y_i \geq 0; \sum_{i=1}^{D} y_i = 1 \right\}.$$

One of the most commonly used probability distributions for compositional data is the Dirichlet distribution, as a Dirichlet distribution of parameter $\alpha \in \mathbb{R}^D$ will generate a D-part compositional vector. Maier (2014) proposed a regression model for Dirichlet-distributed

data [265], which is implemented in the R package *DirichletReg*. However, this model does not take spatial dependencies into account, limiting its applicability in spatial problems.

Over the past decades, spatial autoregressive (SAR) models have emerged as powerful tools for analyzing spatially correlated data in various fields, including economics [35], ecology [106], and epidemiology [167], [213], [345]. The fundamental idea behind SAR models is that the value of a variable at a particular location is influenced not only by its own characteristics but also by the characteristics of neighboring locations. These models explicitly account for the spatial interdependencies among the observed variables, allowing for a more comprehensive understanding of the underlying spatial processes. While spatial dependencies are often present in compositional data, particularly when the components represent different land uses or ecological variables, only a few studies have developed a SAR model for such data. In these studies, the authors employed either a Bayesian estimation approach to estimate the parameters of a spatial multinomial logit model [227] or transform the data into the Euclidian space before applying a multivariate regression model [301], a Gaussian Markov random field [324], or a multivariate conditionally autoregressive model [236].

In order to incorporate spatial dependencies between observations in the case of compositional data, our paper proposes a spatial autoregressive model for Dirichlet-distributed data. We develop a maximum likelihood estimator that effectively handles the spatial interdependency and demonstrate the effectiveness of our model on one synthetic dataset and two real-world datasets.

First, the different models are described in Section 4.2. Then, in Section 5.3, the performances of the models are assessed through synthetic data and three case studies. A discussion is presented in Section 5.4. Finally, we conclude the paper in Section 4.5.

4.2 Methodology

We consider the case where the labels of the dataset are compositional. In this section and the following, let K be the number of features, J the number of classes, n the sample size of the dataset. For the sample i, we define its features as $x_i \in \mathbb{R}^K$ and its label $y_i \in S^J$ (i.e., y_i is an element of the simplex of dimension J). The features (resp., labels) of the whole dataset are then denoted by $X \in \mathbb{R}^{n\times K}$ (resp., $Y \in \mathbb{R}^{n\times J}$). If for a given data row i, the label y_i follows a Dirichlet distribution of parameter $\alpha_i \in \mathbb{R}^J$, then the probability density function is

$$f(y_i|\alpha_i) = \frac{\Gamma(\sum_{j=1}^{J} \alpha_{ij})}{\prod_{j=1}^{J} \Gamma(\alpha_{ij})} \prod_{j=1}^{J} y_{ij}^{\alpha_{ij}-1},$$

where Γ is the gamma function, α_i is called the concentration parameter and has to meet the requirement that $\alpha_{ij} > 0$ for every class j.

The parameters α_i can be parametrized by $\alpha_i = \phi_i \mu_i$ where $\phi_i \in \mathbb{R}$ is the precision parameter (or dispersion parameter) and the compositional vector $\mu_i \in S^J$ represents the individual expected values. Hence, the model's predictions $\hat{y}_i$ is given by the estimated values $\hat{\mu}_i$.

Then, all the parameters α_i can be stacked into a matrix $\alpha \in \mathbb{R}^{n\times J}$. Similarly, we can stack the μ_i (resp. ϕ_i) in a matrix $\mu \in \mathbb{R}^{n\times J}$ (resp. a vector $\phi \in \mathbb{R}^n$).

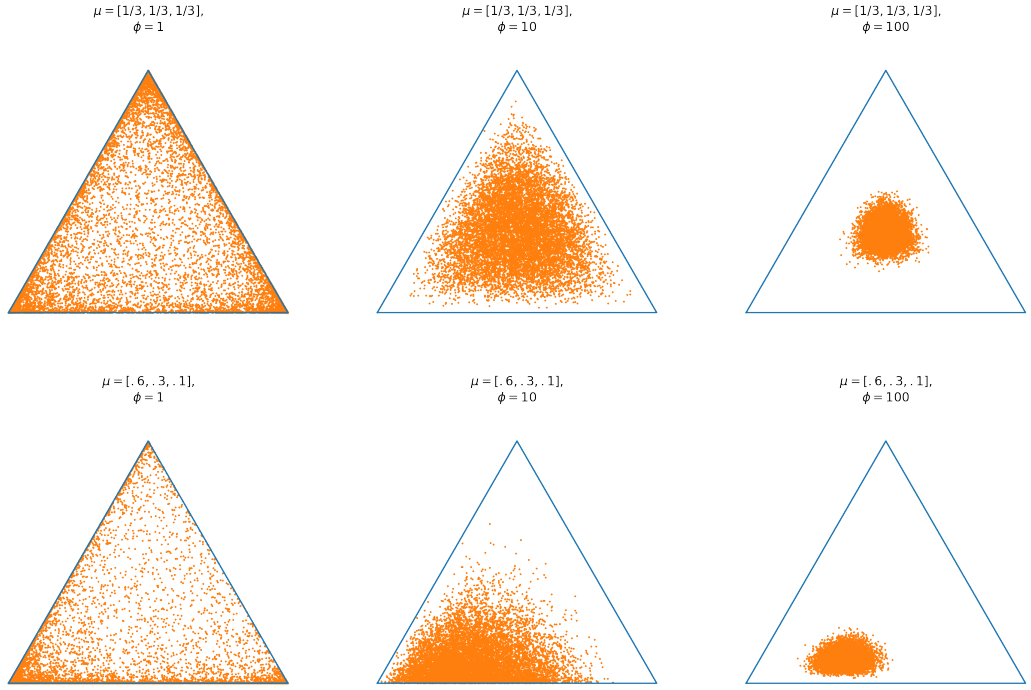


Figure 4.1: Distribution of 10000 points drawn from a Dirichlet distribution, for different values of $\mu \in S^3$ and $\phi \in \mathbb{R}$.

The dispersion parameter ϕ_i plays an important role in the distinction of the classes. For a given μ_i, the smaller ϕ_i is, the more likely the point will be distributed around extreme values (the edges of the simplex). On the contrary, with a high ϕ_i, the point is more likely distributed close to the value of μ_i. This effect is displayed in Figure 4.1, where 10000 points were drawn on a Dirichlet distribution with different parameters ϕ and μ.

When ϕ_i varies, the mean of the points remains the same (μ_i). However, in the case of our regression model, where each point is drawn from a Dirichlet with different parameters, a low ϕ_i would imply that the only drawn point may likely be drawn far from the actual value of μ_i. Hence, in our case, a high value of ϕ_i is preferred. The estimated value of $\hat{\phi}_i$ may also be a good indicator of how accurate we expect the predicted value $\hat{\mu}_i$ to be.

4.2.1 Maximum likelihood regression without spatial lag

Let $\beta \in \mathbb{R}^{K \times J}$ be a matrix of coefficients. We define the $\mu \in \mathbb{R}^{n \times J}$ as

$$\forall i \in [1, \dots, n],\ \forall j \in [1, \dots, J], \quad \mu_{ij} = \frac{\exp(\sum_{k=1}^{K} X_{ik}\beta_{kj})}{\sum_{j'=1}^{J} \exp(\sum_{k=1}^{K} X_{ik}\beta_{kj'})}. \tag{4.1}$$

Then, let $K_Z \in \mathbb{N}$ where $\mathbb{N}$ denotes the set of nonnegative integers. We introduce a matrix $Z \in \mathbb{R}^{n \times K_Z}$ and a vector $\gamma \in \mathbb{R}^{K_Z}$, that allow to define the precision parameter vector $\phi \in \mathbb{R}^n$ as

$$\forall i \in [1, \dots, n], \quad \phi_i = \exp([Z\gamma]_i).$$

For any row i and class j, we then set $\alpha_{ij=}\phi_i\mu_{ij}$, so that $\phi_i = \sum_j \alpha_{ij}$. This parametrization is referred to as the *alternative* parametrization in Maier's paper [265], contrasting it with the *common* parametrization where each ϕ_i is set to 1.

To ensure the unicity of the solution when maximizing the likelihood, the mapping $\beta \mapsto \mu$ must be injective. Because of that, we have to set a column of β as 0, for instance the first column, as done in [265]. The density function can be rewritten depending on μ and ϕ,

$$f(y_i|\mu_i, \phi_i) = \frac{\Gamma(\phi_i)}{\prod_{j=1}^{J} \Gamma(\phi_i \mu_{ij})} \prod_{j=1}^{J} y_{ij}^{\phi_i \mu_{ij} - 1}. \tag{4.2}$$

Thus, the log-likelihood of the Dirichlet distribution is,

$$\ell(y|\mu, \phi) = \sum_{i=1}^{n} \left(\ln \Gamma(\phi_i) - \sum_{j=1}^{J} \ln(\Gamma(\phi_i \mu_{ij})) + \sum_{j=1}^{J} ((\phi_i \mu_{ij} - 1) \ln(y_{ij})) \right) \tag{4.3}$$

$$= \sum_{i=1}^{n} \left(\ln \Gamma(\phi_i) - \sum_{j=1}^{J} \ln\left(\Gamma\left(\phi_i \frac{\exp([X\beta]_{ij})}{\sum_{j'=1}^{J} \exp([X\beta]_{ij'})} \right) \right) + \sum_{j=1}^{J} \left(\left(\phi_i \frac{\exp([X\beta]_{ij})}{\sum_{j'=1}^{J} \exp([X\beta]_{ij'})} - 1 \right) \ln(y_{ij}) \right) \right). \tag{4.4}$$

The presence of the $\ln(y_{ij})$ term in this likelihood function requires y_{ij} to be strictly positive for all i, j. To address the issue of zero values in the data, a possible transformation is to use $y^* = \frac{y(n-1)+1/J}{n}$ [265], which ensures that the transformed values are positive and has the property that $\lim_{n \to +\infty} y^* = y$. For the rest of this paper, we still denote the data as y and assume it does not contain any zero values, but note that the transformation can be applied if necessary.

Because μ and ϕ are parameterized by β and γ, maximum likelihood estimators $\hat{\beta}$ and $\hat{\gamma}$ are used to estimate the parameters β and γ, respectively. To perform second order optimization, and also to obtain the covariance matrix, the gradient and hessian matrix are computed. The details of the computation can be found in Appendix 4.A. Let ψ be the digamma function, i.e., the derivative of the logarithm of the gamma function:

$$\forall x \in \mathbb{R},\ \psi(x) = \frac{\partial}{\partial x} \ln \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}. \tag{4.5}$$

For all $(p, d) \in [1, \ldots, K] \times [1, \ldots, J]$, we have,

$$\frac{\partial}{\partial \beta_{pd}} \ell(y|\mu, \phi) = \sum_{i=1}^{n} \left(\phi_i X_{ip} \mu_{id} \left(\sum_{j=1}^{J} \mu_{ij} \big(\psi(\phi_i \mu_{ij}) - \ln(y_{ij}) \big) - \psi(\phi_i \mu_{id}) + \ln(y_{id}) \right) \right), \tag{4.6}$$

and for $k \in [1, \ldots, K_Z]$,

$$\frac{\partial}{\partial \gamma_k} \ell(y|\mu, \phi) = \sum_{i=1}^{n} Z_{ik} \phi_i \left(\psi(\phi_i) + \sum_{j=1}^{J} \mu_{ij} \big(\ln(y_{ij}) - \psi(\phi_i \mu_{ij}) \big) \right). \tag{4.7}$$

In order to keep the article concise, the expressions of the Hessian matrix are provided only in Appendix 4.A.

With the estimated parameters $\hat{\beta}$ and $\hat{\gamma}$, we are able to predict the label of an unseen data point $\tilde{x} \in \mathbb{R}^K$. This prediction is the compositional vector $\tilde{\mu} \in S^J$, computed from (4.1). The probability vector $\tilde{\mu}$ is considered as being the estimated value of the label.

4.2.2 Maximum likelihood regression with spatial lag

To take into account spatial effect in the model, we introduce a *spatial lag* through the matrix $M = I_n - \rho W$, where I_n is the identity matrix of size n, $\rho \in \mathbb{R}$ is the strength

of spatial correlation and $W \in \mathbb{R}^{n\times n}$ is the spatial weights matrix [22]. It is common to apply row-normalization on W (i.e., its rows sum to 1) [21], which implies that it is often asymmetric even if the original non-normalized weights matrix was symmetric.

Given a matrix of coefficients $\beta \in \mathbb{R}^{K\times J}$, we take $E_i^\Sigma = \sum_{j'=1}^{J} \exp(\sum_{i'=1}^{n}\sum_{k=K}^{n} M_{ii'}^{-1} X_{i'k}\beta_{kj'})$ and redefine μ as:

$$\forall (i,j), \quad \mu_{ij} = \frac{\exp([M^{-1}X\beta]_{ij})}{\sum_{j'=1}^{J}\exp([M^{-1}X\beta]_{ij'})} = \frac{\exp(\sum_{i'=1}^{n}\sum_{k=1}^{K} M_{ii'}^{-1}X_{i'k}\beta_{kj})}{E_i^\Sigma}. \tag{4.8}$$

The introduction of the matrix M modifies the computation of the vector μ to take into account the spatial lag effect in the model. Multiplying the product $X\beta$ with the inverse of M allows to introduce the explanatory variables of the neighboring observations. The value of the spatial correlation parameter ρ needs to be estimated from the data, while W is fixed and has to be defined beforehand. Common choices include distance-based weights or contiguity-based weights [85], [306]. In distance-based weights, the weight between each pair of points is determined by the inverse of the distance between them. In other words, points that are closer to each other receive higher weights, while points that are farther apart receive lower weights. In contiguity-based weights, for each points, the same weight is given to each of its nearest neighbours. The exact criteria for defining neighbors varies depending on the specific application and context, but it generally involves defining adjacency based on some spatial relationship between points. In our case, the nearest neighbours are defined as the closest points as regard to the Euclidian distance.

Let us introduce the matrix $\tilde{X} \in \mathbb{R}^{n\times K}$, such that $\tilde{X} = M^{-1}X$. The loglikelihood remains the same as in (4.4), provided that we replace the term X with $\tilde{X}$ in its expression. Similarly, the calculations of all the gradient and Hessian terms (w.r.t. β and γ) of the loglikelihood are exactly similar to the ones without spatial lag, providing that we replace X by $\tilde{X}$. More specifically, we just have to replace X_{ip} in (4.6) by $\tilde{X}_{ip} = \sum_{i'} M_{ii'}^{-1}X_{i'p}$. Thus,

$$\frac{\partial}{\partial\beta_{pd}}\ell(y|\mu,\phi,\rho) = \sum_{i=1}^{n}\left(\phi_i\tilde{X}_{ip}\mu_{id}\Big(\sum_{j=1}^{J}\mu_{ij}\big(\psi(\phi_i\mu_{ij}) - \ln(y_{ij})\big) - \psi(\phi_i\mu_{id}) + \ln(y_{id})\Big)\right). \tag{4.9}$$

With this model, we also have to compute the derivative with respect to ρ. By defining $U = M^{-1}WM^{-1}X\beta$, which is equal to the derivative of $M^{-1}X\beta$ with respect to ρ, we have

$$\frac{\partial}{\partial\rho}\ell(y|\mu,\phi,\rho) = \sum_{i=1}^{n}\phi_i\sum_{j=1}^{J}\mu_{ij}\Big(\ln(y_{ij})\Big(U_{ij} - \sum_{j'}\mu_{ij'}U_{ij'}\Big) - U_{ij}\Big(\psi(\phi_i\mu_{ij}) - \sum_{j'}\mu_{ij'}\psi(\phi_i\mu_{ij'})\Big)\Big). \tag{4.10}$$

Similar to the previous subsection, to keep the article concise, the detailed computations of the Hessian can be found in Appendix 4.B.

4.2.3 Multinomial distribution and cross-entropy

Let define n independent random variables $\tilde{Y}_i = (\tilde{Y}_{i1}, \ldots, \tilde{Y}_{iJ})$ following a multinomial distribution of probabilities $p_{i1}, \ldots, p_{iJ}$ with n_i trials, denoted by $\tilde{Y}_i \sim \text{Mult}(n_i, p_{i1}, \ldots, p_{iJ})$, where $\sum_j p_{ij} = 1$ and $\sum_j \tilde{Y}_{ij} = n_i$ with $\tilde{Y}_{ij}$ non negative integers. The probability mass function is given in Appendix 4.C.

In order to have the data belonging in the simplex S^J, we can see the y_{ij} as the ratio $\tilde{Y}_{ij}/n_i$. Further, in the same spirit as the Dirichlet regression models, we can define the

multinomial (respectively the spatial multinomial) regression model for compositional data as in (4.1) (resp. (4.8)), using

$$\forall i, j,\ p_{ij} = \frac{\exp([X\beta]_{ij})}{\sum_{j'=1}^{J} \exp([X\beta]_{ij'})} \quad \left(\text{resp. } p_{ij} = \frac{\exp([M^{-1}X\beta]_{ij})}{\sum_{j'=1}^{J} \exp([M^{-1}X\beta]_{ij'})}\right).$$

Here, the model's predictions $\hat{y}_i = (\hat{y}_{i1}, \ldots, \hat{y}_{iJ})$ are given by $\hat{p}_i$, where $\hat{p}_i = (\hat{p}_{i1}, \ldots, \hat{p}_{iJ})$. For the seek of simplicity, we will refer to these models as multinomial regression model (resp. spatial multinomial regression model).

Besides, the cross-entropy between two probability vectors y_i and $\hat{y}_i$ is given by

$$CE_i(y_i, \hat{y}_i) = -\sum_{j=1}^{J} y_{ij} \ln(\hat{y}_{ij}), \tag{4.11}$$

then the cross-entropy loss over the whole dataset is the summation of the CE_i.

The regression parameters of the Dirichlet and multinomial regression models can be estimated either by maximizing the likelihood or minimizing the cross-entropy loss.

We found that minimizing this cross-entropy function sometimes gave a better estimation of the parameters than maximization of the likelihood of the Dirichlet. This behaviour can be explained by the fact that the minimization of the cross-entropy is equivalent to the maximization of the likelihood of a multinomial distribution, given that all the n_i are equal; a detailed description is provided in Appendix 4.C. Hence, we can imagine that when the minimization of the cross-entropy gives better results than the maximization of the likelihood of the Dirichlet, it means the distribution that best fits the data is a multinomial, but not the Dirichlet. This might come from the origin of the compositional data in the datasets we used: they do not naturally represent proportions, but rather a count for different classes (e.g. number of pixels, number of votes) which are then divided by the sum accross all the classes. Because of this, the multinomial distribution is probably more suited to this kind of data.

The minimization of the cross-entropy loss function, for a model with or without spatial parameter, can be compared with the maximization of the likelihood of the Dirichlet in terms of accuracy and computational time. Due to its simpler expression, and probably because it does not need the computation of gamma and polygamma functions, the cross-entropy is faster to optimize.

In the literature, various estimation strategies have been proposed to estimate the parameters of the spatial multinomial distribution (assuming same number of trials n_i for each observation), including a maximum likelihood estimator [70], [243] and a Bayesian estimation strategy [227].

4.2.4 Metrics

In this section, the true probability of class j for sample i is represented by y_{ij}, and the estimated value by the model is denoted as $\hat{y}_{ij}$.

To evaluate and compare the performance of the spatial model and the non-spatial model, several metrics are employed. One commonly used such metric is the Akaike information criterion (AIC). By defining k as the number of estimated parameters and $\hat{\ell}$ as the maximized log-likelihood defined in (4.3), the AIC is calculated as

$$\text{AIC} = -2\hat{\ell} + 2k.$$

Smaller the values of AIC, the better the model's performance.

Another popular metric that used in practice is the cross-entropy, given by

$$\text{Cross-entropy} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} y_{ij}\log(\hat{y}_{ij}).$$

Similar to AIC, smaller the values of cross-entropy, the better the model's performance.

Additionally, the Root Mean Squared Error (RMSE) is utilized to measure model accuracy. As we are dealing with multi-class compositional vectors, the RMSE between a true vector and an estimated vector is computed as the average RMSE across all classes. Specifically, it is calculated using the following formula:

$$\text{RMSE} = \sqrt{\frac{1}{n}\frac{1}{J}\sum_{i=1}^{n}\sum_{j=1}^{J}(y_{ij}-\hat{y}_{ij})^2}.$$

Again, smaller values of the RMSE indicate better model performance.

Furthermore, the coefficient of determination, typically denoted as R^2, can be computed on each class. For a given class j,

$$R_j^2 = 1 - \frac{\sum_{i=1}^{n}(y_{ij}-\hat{y}_{ij})^2}{\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)^2}, \tag{4.12}$$

where $\bar{y}_j$ represents the mean of the values in the jth class. The overall R^2 value is obtained by averaging R_j^2 values across all classes. Note that, although it may seem counter-intuitive, this definition allows for R^2 to be negative in cases where the predicted values $\hat{y}_{ij}$ deviate more from the true values than the mean $\bar{y}_j$. It can happen in the cases when the true values are tightly clustered together (which can easily happen with compositional data), their deviation from the mean becoming very small, making the numerator in (4.12) close to zero. In such cases, the R^2 value tends to be negative or approach negative infinity.

Finally, the metric that is probably the most suitable to compare the distance between two vectors labels is the cosine similarity, representing the cosine of the angle formed between the two vectors, given by

$$\text{Cos similarity} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i\hat{y}_i^T}{\|y_i\|_2\cdot\|\hat{y}_i\|_2}.$$

4.3 Results

In the following sections, we present the results obtained from applying both the spatial lag model and the non-spatial model to various datasets. Each dataset is described in detail, along with the corresponding outcomes achieved by the respective models.

4.3.1 Synthetic dataset

The synthetic spatially-correlated dataset is generated with 2 features and 3 classes, by varying the number of samples n (50, 200, or 1000) and the values of ρ (0.1, 0.5, or 0.9), which are all positive, as most often found in literature [23]. The values of β and γ are predetermined at the beginning of the simulation. In the presented results, the parameters

are

$$\beta = \begin{matrix} \\ \text{features} \end{matrix} \overset{\text{classes}}{\begin{bmatrix} 0 & 0 & 0.1 \\ 0 & 1 & -2 \\ 0 & -1 & -2 \end{bmatrix}}, \qquad \gamma = \begin{bmatrix} 2 \\ 3 \end{bmatrix}.$$

These specific values were selected to achieve certain desirable properties in the generated synthetic data. In particular, the value of β has been chosen to ensure that the classes are balanced in μ, i.e., that no class is significantly more frequent or rarer than others. With such a value for β, the data are generated with class distributions that are evenly spread, avoiding any class dominance or extreme imbalances. Regarding γ, its chosen value was intended to ensure that the precision parameter ϕ, which controls the spread or dispersion of the class probabilities, is sufficiently high, so that the distribution of the points is relatively concentrated around their class probabilities. This way, the class patterns are more distinguishable and well-defined.

Initially, n samples are randomly drawn from a multivariate normal distribution with two covariates, producing the features matrix $X \in \mathbb{R}^{n\times 2}$. The matrix Z is constructed with one covariate drawn from a uniform distribution. To build the matrix W, we assign its row index to each sample and identify its nearest neighbors based on these indices. In our simulation, we specifically considered 5 neighbors, but additional simulations involving more neighbors suggested comparable results.

An alternative approach, based on [47], [68], is to construct the matrix W by selecting a radius of a specific size (depending on the value of ρ) around each point in X, and to consider all the values falling within this radius as neighbors of the corresponding point. This method, however, did not seem to recreate the behaviour of a spatial effect, as it only repeated the information already present in the matrix X.

Then, from the matrices X and Z and the parameters β and γ, we compute μ and ϕ that are then used to produce α. The response matrix Y is finally generated by drawing in the Dirichlet distribution of parameter α_i for each row i.

We repeat 100 times the following experiment: the data are created and the bias of the estimated parameters is computed. The results for each of the three ρ are presented in Tables 4.1, 4.2 and 4.3.

In regard to the non-spatial parameters, both models demonstrate similar behavior, with their bias, variance, and mean squared error being asymptotically unbiased. In the spatial models, we also observe the expected behavior, where the bias and mean squared error of the estimated $\hat{\rho}$ decrease as the number of samples increases. It is worth noting that when $\hat{\rho}$ is biased (which occurs when the sample size n is small), the bias is negative, suggesting that the model tends to underestimate the spatial correlation strength.

The prediction accuracy of the models is assessed as follow. For each value of ρ, we create the test set by generating 1000 new data points using the true parameters β^*. On this test data, the true value μ^* is computed. Then, we compute $\hat{\mu}$ using the parameters $\hat{\beta}$ estimated from the $n = 1000$ simulation. To evaluate the difference between μ^* and each $\hat{\mu}$, metrics such as R^2, RMSE, cross-entropy and cosine similarity are utilized. The results are presented in Table 4.4. For a low spatial correlation ($\rho = 0.1$), both models perform equally well. However, as the spatial correlation increases ($\rho = 0.5$ and $\rho = 0.9$), the performances of the non-spatial models decrease, and they are outperformed by the spatial model across all metrics. Interestingly, the spatial model has best performance when the spatial correlation

Table 4.1: Bias, standard deviation and mean squared error of the estimated parameters ($\rho = 0.1$). The results are presented as the mean of the differences on the 100 iterations, the standard deviation within parenthesis, and the mean squared error within square brackets.

	Spatial			Not spatial		
Parameter	$n = 50$	$n = 200$	$n = 1000$	$n = 50$	$n = 200$	$n = 1000$
β_{01}	0.013 (0.059) [0.004]	0.012 (0.032) [0.001]	0.010 (0.016) [0.0]	0.014 (0.067) [0.005]	0.012 (0.036) [0.001]	0.011 (0.018) [0.0]
β_{02}	0.038 (0.087) [0.009]	0.035 (0.034) [0.002]	0.031 (0.016) [0.001]	0.051 (0.106) [0.014]	0.055 (0.046) [0.005]	0.047 (0.021) [0.003]
β_{11}	-0.062 (0.079) [0.010]	-0.037 (0.038) [0.003]	-0.019 (0.017) [0.001]	-0.067 (0.081) [0.011]	-0.040 (0.038) [0.003]	-0.021 (0.017) [0.001]
β_{12}	0.277 (0.12) [0.091]	0.185 (0.051) [0.037]	0.141 (0.023) [0.020]	0.288 (0.127) [0.099]	0.198 (0.055) [0.042]	0.158 (0.024) [0.026]
β_{21}	0.042 (0.083) [0.009]	0.015 (0.035) [0.001]	0.004 (0.014) [0.0]	0.045 (0.081) [0.009]	0.017 (0.035) [0.002]	0.007 (0.015) [0.0]
β_{22}	0.197 (0.108) [0.051]	0.123 (0.05) [0.018]	0.088 (0.023) [0.008]	0.203 (0.105) [0.052]	0.131 (0.053) [0.02]	0.101 (0.025) [0.011]
γ_0	0.534 (0.312) [0.382]	0.353 (0.139) [0.144]	0.269 (0.066) [0.077]	0.522 (0.32) [0.375]	0.349 (0.144) [0.143]	0.264 (0.066) [0.074]
γ_1	-0.354 (0.598) [0.483]	-0.324 (0.246) [0.165]	-0.294 (0.105) [0.098]	-0.453 (0.612) [0.58]	-0.448 (0.266) [0.272]	-0.416 (0.11) [0.185]
ρ	-0.01 (0.052) [0.003]	-0.003 (0.023) [0.001]	-0.0 (0.009) [0.0]	/ /	/ /	/ /

Table 4.2: Bias, standard deviation and mean squared error of the estimated parameters ($\rho = 0.5$). The results are presented as the mean of the differences on the 100 iterations, the standard deviation within parenthesis, and the mean squared error within square brackets.

	Spatial			Not spatial		
Parameter	$n = 50$	$n = 200$	$n = 1000$	$n = 50$	$n = 200$	$n = 1000$
β_{01}	0.009 (0.042) [0.002]	0.011 (0.02) [0.001]	0.007 (0.009) [0.0]	0.029 (0.157) [0.026]	0.03 (0.079) [0.007]	0.028 (0.037) [0.002]
β_{02}	0.012 (0.047) [0.002]	0.023 (0.017) [0.001]	0.017 (0.009) [0.0]	0.116 (0.312) [0.111]	0.184 (0.147) [0.055]	0.144 (0.066) [0.025]
β_{11}	-0.06 (0.083) [0.011]	-0.037 (0.037) [0.003]	-0.024 (0.016) [0.001]	-0.163 (0.122) [0.042]	-0.17 (0.057) [0.032]	-0.155 (0.027) [0.025]
β_{12}	0.356 (0.123) [0.142]	0.221 (0.061) [0.052]	0.176 (0.029) [0.032]	0.682 (0.205) [0.507]	0.622 (0.099) [0.397]	0.598 (0.049) [0.36]
β_{21}	0.051 (0.067) [0.007]	0.016 (0.041) [0.002]	0.01 (0.015) [0.0]	0.142 (0.123) [0.035]	0.129 (0.055) [0.02]	0.129 (0.028) [0.017]
β_{22}	0.246 (0.108) [0.072]	0.145 (0.049) [0.023]	0.113 (0.022) [0.013]	0.511 (0.202) [0.302]	0.472 (0.088) [0.23]	0.462 (0.042) [0.216]
γ_0	0.523 (0.314) [0.372]	0.418 (0.153) [0.198]	0.285 (0.062) [0.085]	0.015 (0.364) [0.132]	-0.178 (0.164) [0.059]	-0.337 (0.087) [0.121]
γ_1	-0.363 (0.637) [0.538]	-0.395 (0.289) [0.24]	-0.313 (0.109) [0.11]	-1.828 (0.749) [3.903]	-2.0 (0.293) [4.086]	-1.92 (0.137) [3.704]
ρ	-0.005 (0.032) [0.001]	-0.002 (0.011) [0.0]	0.0 (0.005) [0.0]	/ /	/ /	/ /

Table 4.3: Bias, standard deviation and mean squared error of the estimated parameters ($\rho = 0.9$). The results are presented as the mean of the differences on the 100 iterations, the standard deviation within parenthesis, and the mean squared error within square brackets.

	Spatial			Not spatial		
Parameter	$n = 50$	$n = 200$	$n = 1000$	$n = 50$	$n = 200$	$n = 1000$
β_{01}	0.007 (0.076) [0.006]	0.011 (0.021) [0.001]	0.01 (0.011) [0.0]	0.096 (0.596) [0.364]	0.019 (0.24) [0.058]	0.044 (0.097) [0.011]
β_{02}	-0.022 (0.126) [0.016]	-0.019 (0.03) [0.001]	-0.024 (0.017) [0.001]	0.303 (1.096) [1.292]	0.298 (0.381) [0.234]	0.206 (0.165) [0.069]
β_{11}	-0.418 (0.335) [0.287]	-0.423 (0.194) [0.217]	-0.39 (0.112) [0.165]	-0.653 (0.254) [0.491]	-0.696 (0.113) [0.496]	-0.673 (0.05) [0.455]
β_{12}	1.193 (0.481) [1.653]	1.185 (0.345) [1.523]	1.148 (0.218) [1.365]	1.588 (0.27) [2.596]	1.559 (0.088) [2.437]	1.531 (0.049) [2.345]
β_{21}	0.356 (0.271) [0.2]	0.368 (0.196) [0.174]	0.359 (0.114) [0.142]	0.611 (0.233) [0.427]	0.655 (0.108) [0.441]	0.652 (0.052) [0.428]
β_{22}	0.997 (0.434) [1.182]	1.003 (0.347) [1.127]	0.976 (0.218) [1.0]	1.444 (0.253) [2.149]	1.446 (0.114) [2.104]	1.444 (0.056) [2.089]
γ_0	0.411 (0.679) [0.629]	-0.257 (0.564) [0.384]	-0.414 (0.419) [0.347]	-1.24 (0.645) [1.952]	-1.924 (0.149) [3.724]	-2.174 (0.072) [4.731]
γ_1	-1.909 (0.759) [4.218]	-2.151 (0.547) [4.927]	-2.167 (0.323) [4.8]	-2.717 (0.474) [7.606]	-2.787 (0.128) [7.784]	-2.779 (0.065) [7.729]
ρ	-0.011 (0.051) [0.003]	-0.006 (0.031) [0.001]	-0.004 (0.017) [0.0]	/ /	/ /	/ /

strength is moderately high ($\rho = 0.5$), exhibiting slightly better results compared to the scenario with an extremely high correlation strength ($\rho = 0.9$). This suggests that the spatial model is most effective when there is a moderate level of spatial dependence, and its performance may plateau or decline at extremely high spatial correlation levels.

Table 4.4: Scores between the μ^* of the test set, and the estimated $\hat{\mu}$ (computed with the parameters estimated with $n = 1000$). The results are displayed as mean on the 100 iterations (standard deviation).

	Model	$\boldsymbol{R^2}$	**RMSE**	**Cross-entropy**	**Cos similarity**
$\rho = 0.1$	Not spatial	0.9309 ($< 10^{-4}$)	0.0739 ($< 10^{-4}$)	0.6660 (0.0001)	0.9855 ($< 10^{-4}$)
	Spatial	0.9335 ($< 10^{-4}$)	0.0723 ($< 10^{-4}$)	0.6648 (0.0001)	0.9862 ($< 10^{-4}$)
$\rho = 0.5$	Not spatial	0.8311 (0.0001)	0.1249 ($< 10^{-4}$)	0.6845 (0.0001)	0.9610 ($< 10^{-4}$)
	Spatial	0.9408 ($< 10^{-4}$)	0.0705 ($< 10^{-4}$)	0.6275 (0.0002)	0.9872 ($< 10^{-4}$)
$\rho = 0.9$	Not spatial	0.2073 (0.0025)	0.3257 (0.0001)	0.9033 (0.0005)	0.7764 (0.0001)
	Spatial	0.9011 (0.0026)	0.1097 (0.0008)	0.4414 (0.0026)	0.9776 (0.0001)

We then also try to retrieve the parameters from a multinomial model. As a reminder, when the number of trials is equal for each observation, the estimation of the multinomial model using the maximum likelihood is equivalent to minimizing the cross-entropy. The results, presented in Appendix 4.D, were similar to those of the Dirichlet model. However, we observed that when the γ parameter was decreased, leading to less distinguishable classes, the performances of the Dirichlet model significantly deteriorated compared to the multinomial model.

Additionally, we conducted experiments where the labels were generated using a multinomial distribution instead of a Dirichlet distribution. In these cases, the performance of the models depended on the distinguishability of the classes, determined by the parameter n_i, representing the number of trials for each observation i. When the n_i values were small and drawn from a discrete uniform distribution between 1 and 100, the multinomial model

outperformed the Dirichlet model. However, with larger n_i values, for instance ranging from 1000 to 10000, both models achieved excellent performances, and their results became comparable. Interestingly, even when each observation was generated with a different n_i, the multinomial model assuming equal sample sizes (i.e. the minimization of cross-entropy) yielded similar performances to the full multinomial model.

4.3.2 Arctic Lake

The Arctic Lake dataset [87] provides compositional data in terms of sand, silt, and clay percentages for 39 sediment samples taken at various water depths in an Arctic lake. The goal is to quantify the extent to which water depth influences the compositional patterns of the sediment samples. We propose here two regression models: one with a single predictor variable (the depth) along with an intercept term, and another model that includes an intercept term, the depth variable, and its squared value as additional predictor.

Due to the limited size of the dataset, we employ a Leave One Out Cross Validation (LOOCV) strategy. We iteratively exclude one sample (the k-th sample) from the dataset, and use the remaining samples to estimate the parameters of the models. We then compute the predicted odd values $\hat{\mu}_k$ for the excluded sample, and evaluate its proximity to the true compositional label using metrics such as R^2, RMSE, cross-entropy and cosine similarity. The mean and standard deviation of these metrics are reported in Table 4.5.

Table 4.5: Scores for the models on Arctic Lake dataset with a LOOCV strategy. The scores are the mean on all the folds, and standard deviation within parenthesis.

Model		R^2	RMSE	Cross-entropy	Cos similarity
Order 1	without spatial	0.5887 (0.015)	0.1015 (0.0018)	0.9134 (0.0027)	0.9665 (0.0012)
	spatial contiguity	**0.6323** (0.0147)	**0.0951** (0.0017)	**0.9064** (0.0029)	**0.9706** (0.0011)
	spatial distance	0.5893 (0.0148)	0.1014 (0.0018)	0.9134 (0.0027)	0.9665 (0.0012)
Order 2	without spatial	0.6784 (0.0144)	0.0881 (0.0019)	0.8993 (0.0032)	0.9743 (0.0011)
	spatial contiguity	**0.6943** (0.0144)	**0.0858** (0.0019)	**0.8974** (0.0033)	**0.9755** (0.0011)
	spatial distance	0.6863 (0.015)	0.0871 (0.002)	0.8982 (0.0032)	0.9747 (0.0012)

The results suggest that the utilization of spatial information leads to slight improvements in model performance. However, in terms of variability, the difference may not be statistically significant. This lack of significance could be attributed to the spatial information being derived solely from the depth variable, resulting in the absence of any new information being introduced. Instead, the data is essentially replicated in a different manner.

4.3.3 Maupiti

Maupiti Island, situated in the Society archipelago of French Polynesia, is an island spanning approximately 8km by 8km in size. The dataset originates from an expert-driven mapping process of Maupiti Island, based on various field observation campaigns [394]. Details on how the dataset has been created can be found on a previous study on compositional data [299]. Each of the 2091 samples of the dataset are segments created via Felzenszwalb's segmentation [117] on the RGB image, each of them having 16 features (4 statistical moments on 4 different bands) and a compositional data label composed of 4 classes: coral, sand, shorereef, and mixed. The label is derived from the expert's ground-truth map: after the segmentation process, we count the pixels assigned to each class within each segment, and divide it by the

total number of pixels within the respective segment. Note that in this case the number of pixels in each segment is different.

The neighbors matrix W is created as a distance-based matrix between each segment and its neighbor. Note that we also experimented with a contiguity-based matrix, but it yielded inferior performances. Since our analysis specifically focuses on the shallow regions of the lagoon, which offer clearer and more easily interpretable imagery, some segments are excluded from the analysis, which in turn can result in certain segments not having any neighbors.

The matrix Z is chosen as being a copy of the features matrix X, as this choice yielded the best results compared to the case where Z is only an intercept.

We use the Dirichlet models to retrieve the parameters and then, an estimated $\hat{\mu}$ is computed and compared with the real μ. The metrics to compare them are reported in Table 4.6. The spatial correlation parameter $\hat{\rho}$ is estimated to be 0.93 with this dataset, indicating a high spatial correlation between these segments.

Table 4.6: Scores for the Dirichlet models on Maupiti dataset.

Model	$\boldsymbol{R^2}$	**RMSE**	**Cross-entropy**	**AIC**	**Cos similarity**
Without spatial	0.265	0.297	0.815	-73873	0.786
With spatial	0.441	0.261	0.675	-74170	0.848

The spatial model consistently outperforms the non-spatial model across all metrics. However, it is worth noting that the performance of the spatial model remains relatively low, with an R^2 value of 0.441 and a cosine similarity of 0.848. This suggests that the use of the Dirichlet distribution may not be well-suited for this specific dataset. It is currently unclear whether this limitation is specific to the unique characteristics of the Maupiti dataset or if the Dirichlet distribution generally lacks suitability for datasets involving geographical maps. Furthermore, Table 4.7 reveals that the minimization of the cross-entropy function on this dataset yields significantly better performances than the Dirichlet regressor, further confirming that the Dirichlet distribution may not be well suited for this dataset.

Table 4.7: Scores for the models on Maupiti dataset using the minimization of the cross-entropy.

Model	$\boldsymbol{R^2}$	**RMSE**	**Cross-entropy**	**Cos similarity**
Without spatial	0.636	0.221	0.455	0.868
With spatial	0.820	0.160	0.307	0.925

Furthermore, we compute the maximum a posteriori (MAP) by taking the argmax of the compositional data label for each segment and assign it as the class of the segment. This process allows for the creation of a visually interpretable map. The ground-truth map resulting from the expert-based mapping is depicted in Figure 4.2a, along with the maps obtained from the predictions of the non-spatial (Figure 4.2b) and spatial Dirichlet models (Figure 4.2c). To evaluate the accuracy of these models, pixel-wise comparisons are conducted to count the number of correctly assigned pixels, which is then divided by the total number of classified pixels, excluding masked pixels. The accuracy achieved by the non-spatial model is 0.752, while the spatial model achieves an accuracy of 0.832, further confirming that the spatial model provides superior results.

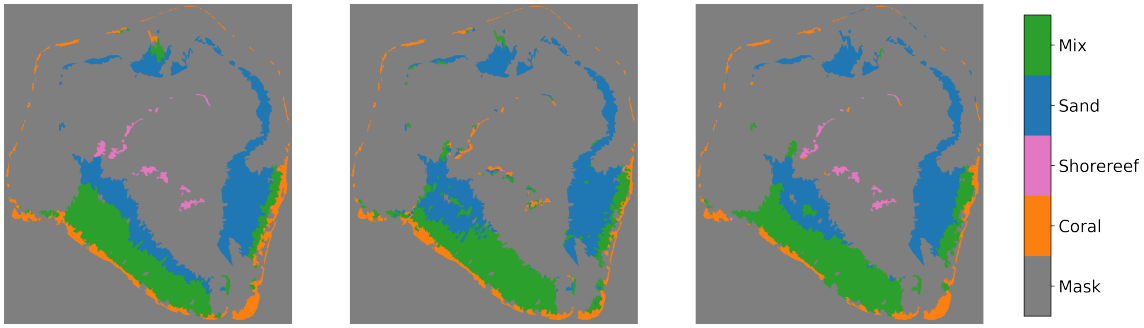


Figure 4.2: Maps created by using (a) the real labels, (b) the labels computed with the non-spatial Dirichlet model, and (c) the labels computed with the spatial Dirichlet model. The pixel-wise accuracy between maps (a) and respectively (b) and (c) is 0.752 and 0.832.

4.3.4 Elections

This dataset present the votes at the French departmental election of 2015 in the Occitanie region [160], [300], for $n = 207$ cantons. For each canton, the voting distribution (initially between 15 political parties) is categorized into three major political movements: left, right, and extreme right. In our study, we utilized 25 distinct social indicators as features, including age categories, employment fields, and education level, among others. Initially, the dataset consisted of 283 cantons, but any cantons where one of the classes was not present were removed. This resulted in the exclusion of 76 points, which accounts for 27% of the data.

The spatial weights matrix W is computed based on the geographic proximity of each canton's center. Specifically, two cases are considered. In the first case, the contiguity-based, we consider the 5 nearest neighboring cantons, determined by their center-to-center distances. In the second case, the distance-based, the inverse of the distance between each canton and the others is considered, with a cut-off at a certain value that minimizes the average number of neighbors and to ensure that each canton has at least one neighbor. This cut-off gives 12 neighbors on average.

The matrix Z is chosen as being a sole intercept, as this choice yielded the best results compared to the case where Z is a copy of the features matrix X.

Then, for the three models (non-spatial and the two spatials), we use the maximum likelihood estimator to retrieve the parameters and compute the performance with our usual metrics. Results are reported in Table 4.8. The estimated spatial correlation coefficient $\hat{\rho}$ is 0.97 (resp. 0.91) with the distance-based (resp. contiguity-based) matrix.

Table 4.8: Scores for the Dirichlet models on Elections dataset.

Model	$\boldsymbol{R^2}$	**RMSE**	**Cross-entropy**	**AIC**	**Cos similarity**
No spatial	0.487	0.080	1.048	-862.1	0.975
Spatial (contiguity)	0.582	0.072	1.042	-947.4	0.979
Spatial (distance)	**0.602**	**0.070**	**1.041**	**-965.1**	**0.980**

The spatial models perform better than the non-spatial model across all evaluation metrics, excepted for the AIC which is slightly better for the non-spatial model. Besides, the distance-based spatial model performs slightly better than the contiguity-based. Additionally, we attempted to make predictions using our model through a 10-fold cross-validation technique. In this approach, 90% of the data were used to estimate the model parameters, while the

remaining 10% (corresponding to 21 values) were reserved for testing the model's performance. However, we observed extremely poor performance on the test set, indicating that the spatial model is highly sensitive to missing values. This could be attributed to the fact that 27% of the initial data was already missing, and further data removal might have rendered the spatial information irrelevant. Notably, in a previous study [160], the authors were able to successfully make predictions using the entire dataset of 283 cantons.

An important observation with this dataset is that the multinomial distribution and the minimization of the cross-entropy yielded similar results compared to the Dirichlet distribution.

4.4 Discussion

First and foremost, it is important to remember that the minimization of cross-entropy is equivalent to maximizing the likelihood of a multinomial distribution when each number of trials is equal. However, even when they are different, the results remain quite similar. We conducted several tests by generating synthetic datasets with varying sample sizes, and interestingly, the performances of a multinomial regressor using sample size information were comparable to those of a regressor assuming similar sample sizes. Thus, we will refer to both cases as the "multinomial model", which encompasses the minimization of cross-entropy. Additionally, when we evaluated a Dirichlet regression model on these multinomial generated data, we found that it performed similarly to the multinomial model, excepted in the cases of extremely high spatial correlation, where the multinomial data was slightly better.

For the Dirichlet regression models, the analysis of the synthetic dataset reveals that both the spatial and non-spatial estimators are asymptotically unbiased. As shown in Table 4.4, when the spatial correlation in the data is low ($\rho = 0.1$), there is no significant difference in the performances between the spatial and non-spatial models. This finding was further confirmed when we generated data without spatial dependencies, and the spatial model accurately retrieved the parameters, estimating that $\hat{\rho}$ was not significantly different from 0. On the other hand, when the spatial correlation is high ($\rho = 0.5$ or 0.9), the spatial model outperforms the non-spatial one. Notably, the spatial model exhibits better performance under moderately high spatial correlation ($\rho = 0.5$) compared to extremely high spatial correlation ($\rho = 0.9$). This behavior may be attributed to the challenge of distinguishing between spatial patterns and the true underlying relationships between the variables when the spatial correlation is exceedingly strong. However, interestingly, under the extremely high spatial correlation scenario, the multinomial model performs slightly better than the Dirichlet model for the prediction task, as observed when comparing Tables 4.4 and 4.12. This suggests that the multinomial regression model may be more suited to handle high spatial correlation scenarios.

Overall, the analysis of the synthetic dataset reveals that the performance of the spatial Dirichlet model is influenced by the level of spatial correlation present in the data. While the spatial model demonstrates improved estimation in cases of spatially correlated data, it still exhibits non-optimal performance when dealing with strong spatial correlations. This behavior could partially explain the suboptimal results of the Dirichlet model on the Maupiti dataset (Table 4.6), where the estimated $\hat{\rho}$ was 0.93. However, as the model performs well on the Elections dataset, with a high estimated $\hat{\rho}$ (> 0.9), other factors may contribute to

the poor performances on Maupiti dataset. Furthermore, the fact that the minimization of the cross-entropy yielded higher performances (Table 4.7) raises questions about whether the Dirichlet distribution was truly suitable for this particular dataset. Additionally, our experiments with the multinomial distribution taking into account the sample sizes, did not outperform the cross-entropy approach when assuming equal sample size. This observation aligns with the findings from the synthetic data analysis mentioned earlier in this section.

From our analysis of the real datasets, several important conclusions can be drawn regarding the impact of spatial information on model performance. We observed that incorporating spatial information can significantly improve results when the spatial information truly represents additional data, rather than being derived from existing data. For instance, in the case of the Arctic Lake dataset, when we attempted to recreate the spatial weight matrix W using only the available covariate (depth), there was no significant improvement in model performance (Section 4.3.2). This emphasizes the importance of incorporating genuinely new spatial information to achieve better results.

Furthermore, when a Dirichlet distribution is well-suited to the dataset, as demonstrated in the Elections dataset (Section 4.3.4), our SAR Dirichlet model outperforms the non-spatial model. Notably, the spatial model exhibits robust performance even in the presence of missing data, with 27% of the initial data missing. However, when a large amount of data is missing, as observed during a 10-fold cross-validation, the model's performance tends to degrade.

In our analysis of real-life datasets, we observed that the spatial weights matrix W performed better when defined as distance-based rather than contiguity-based. However, it is essential to acknowledge that this result may not be generalized to every scenario, and it might be specific to the datasets we examined. We have not been able to find comprehensive studies in literature providing a definitive analysis to determine the best type of spatial weights matrix for all cases.

In the presented work, we estimated the parameters of the models through a probabilistic approach. This approach offers distinct advantages, primarily by providing quantifiable measures of uncertainty, including p-values, confidence intervals, and enabling statistical inference. To evaluate the significance of the spatial parameter ρ, we conducted Wald tests and Log-ratio tests (LRT). However, our analysis revealed that in the context of our SAR Dirichket model, these tests did not effectively control the significance level α. Furthermore, we extended our investigation to encompass general linear SAR models and encountered similar issues with controlling the significance level for the spatial parameter ρ. Notably, the tests applied did not maintain the desired α level. It's worth noting that our exploration into the literature did not yield studies that have extensively examined this phenomenon, leaving us uncertain about whether the observed effects are due to a potential oversight in our methodology or if they reflect broader trends in SAR models. Further research and investigation are necessary to shed light on this matter.

4.5 Conclusion

Our study demonstrates that incorporating spatial dependencies in a Dirichlet model leads to improved performance when dealing with datasets featuring compositional labels. Our findings from the real-life datasets reveal that a distance-based spatial weight matrix tends

to yield better results compared to a contiguity-based matrix. These results underscore the potential advantages of spatial modeling, especially in scenarios where the Dirichlet distribution is well-suited to the data.

The results obtained from the synthetic dataset provide some insights into the behavior of the SAR Dirichlet model. While the spatial model outperforms the non-spatial model in spatially correlated data, the spatial model does not perform optimally under extremely high spatial correlation and provides better results when the spatial correlation is moderate ($\rho = 0.5$).

Overall, our study highlights the importance of considering spatial information when it provides meaningful additional context, as it can significantly enhance the model's effectiveness. It also emphasizes the potential impact of missing data, which should be carefully addressed to avoid adverse effects on model performance.

Moreover, our findings suggest that, in general, the multinomial model appears to be better suited for handling compositional data compared to the Dirichlet model, especially in the cases where the class patterns are not highly distinguishable. Throughout our analysis, we did not encounter instances where the Dirichlet model significantly outperformed the multinomial model. However, we believe that this behaviour may be caused by the fact that the compositional data present in our datasets are not naturally a probability, but a ratio (the count of the number of pixels, or the number of votes). The multinomial distribution might be naturaly more suited to handle this case. Further investigations are required to determine if it holds true across a broader range of scenarios and datasets.

Appendix 4.A Computation on Dirichlet Distribution without Spatial Lag

Let K be the number of features, J be the number of classes and n be the number of samples. We have the features matrix $X \in \mathbb{R}^{n\times K}$, the label matrix $Y \in \mathbb{R}^{n\times J}$ and the matrix of parameters $\alpha \in \mathbb{R}^{n\times J}$. To make the computations easier, we split the loglikelihood into three parts A_i, B_i and C_i as follows.

$$\ell(y|\mu,\phi) = \sum_{i=1}^{n} \left(\underbrace{\ln\Gamma(\phi_i)}_{A_i} - \underbrace{\sum_j \ln(\Gamma(\phi_i\mu_{ij}))}_{B_i} + \underbrace{\sum_j ((\phi_i\mu_{ij}-1)\ln(y_{ij}))}_{C_i} \right). \tag{4.13}$$

Note that μ is a function of β and ϕ is a function of γ. Below, in Sections 4.A.1 and 4.A.2 we respectively provide computations of the first order and the second order derivatives of the loglikelihood function with respect to β and γ.

4.A.1 First order derivative

To compute the gradient of the loglikelihood, we compute the derivatives of A_i, B_i and C_i. To do so, we first compute the derivative of μ_{ij} with respect to β_{pd} for a given feature p and a given class d. Let

$$E_i^{\Sigma} = \sum_{j'=1}^{J} \exp(\sum_{k=1}^{K} X_{ik}\beta_{kj'}).$$

We consider the derivatives in two seperate cases.

- **Case $j \neq d$:**

$$\begin{aligned} \frac{\partial\mu_{ij}}{\partial\beta_{pd}} &= \frac{-\exp(\sum_k X_{ik}\beta_{kj}) \times X_{ip}\exp(\sum_k X_{ik}\beta_{kd})}{(E_i^{\Sigma})^2} \\ &= -X_{ip}\mu_{ij}\mu_{id} \end{aligned} \tag{4.14}$$

- **Case $j = d$:**

$$\begin{aligned} \frac{\partial\mu_{id}}{\partial\beta_{pd}} &= \frac{X_{ip}\exp(\sum_k X_{ik}\beta_{kd}) \times E_i^{\Sigma} - \exp(\sum_k X_{ik}\beta_{kd}) \times X_{ip}\exp(\sum_k X_{ik}\beta_{kd})}{(E_i^{\Sigma})^2} \\ &= X_{ip}\mu_{id} - X_{ip}(\mu_{id})^2 \\ &= X_{ip}\mu_{id}(1-\mu_{id}) \end{aligned} \tag{4.15}$$

Since ϕ_i is not a function of β, for all p and d, we have $\frac{\partial A_i}{\partial\beta_{pd}} = 0$ and

$$\frac{\partial\phi_i\mu_{ij}}{\partial\beta_{pd}} = \phi_i\frac{\partial\mu_{ij}}{\partial\beta_{pd}}.$$

Now let ψ be the digamma function, defined in (4.5). From (4.14) and (4.15), the derivatives of B_i and C_i are

$$\frac{\partial B_i}{\partial\beta_{pd}} = \phi X_{ip}\mu_{id}(1-\mu_{id}) \cdot \frac{\partial}{\partial\phi_i\mu_{id}} \ln\big(\Gamma(\phi_i\mu_{id})\big) - \phi_i \sum_{j\neq d} X_{ip}\mu_{ij}\mu_{id} \cdot \frac{\partial}{\partial\phi_i\mu_{ij}} \ln\big(\Gamma(\phi_i\mu_{ij})\big)$$

$$= \phi_i X_{ip}\mu_{id}\psi(\phi_i\mu_{id}) - \phi_i \sum_j X_{ip}\mu_{ij}\mu_{id}\psi(\phi_i\mu_{ij})$$

$$= \phi_i X_{ip}\mu_{id}\left(\psi(\phi_i\mu_{id}) - \sum_j \mu_{ij}\psi(\phi_i\mu_{ij})\right), \tag{4.16}$$

and

$$\frac{\partial C_i}{\partial \beta_{pd}} = X_{ip}\mu_{id}(1-\mu_{id})\cdot\frac{\partial}{\partial \mu_{id}}(\phi_i\mu_{id}-1)\ln(y_{id}) - \sum_{j\neq d} X_{ip}\mu_{ij}\mu_{id}\cdot\frac{\partial}{\partial \mu_{ij}}(\phi_i\mu_{ij}-1)\ln(y_{ij})$$

$$= \phi_i X_{ip}\mu_{id}\left(\ln(y_{id}) - \sum_j \mu_{ij}\ln(y_{ij})\right). \tag{4.17}$$

Summing the results of (4.16) and (4.17) gives (4.6).

We now compute the derivatives of the loglikelihood with respect to γ_k for $k \in [1,\ldots,K_Z]$. Since

$$\frac{\partial \phi_i}{\partial \gamma_k} = Z_{ik}\phi_i,$$

using the change of variables,

$$\begin{aligned}
\frac{\partial A_i}{\partial \gamma_k} &= Z_{ik}\phi_i\psi(\phi_i),\\
\frac{\partial B_i}{\partial \gamma_k} &= \sum_j \psi(\phi_i\mu_{ij})\mu_{ij}\frac{\partial \phi_i}{\partial \gamma_k} = Z_{ik}\phi_i\sum_j \mu_{ij}\psi(\phi_i\mu_{ij}), \quad \text{and}\\
\frac{\partial C_i}{\partial \gamma_k} &= Z_{ik}\phi_i\sum_j \mu_{ij}\ln(y_{ij}).
\end{aligned}$$

Thus, the summation of the above three expression results as (4.7).

4.A.2 Second order derivative

The Hessian matrix of the loglikelihhod function (4.13) is the collection of derivatives of (4.6) with respect to all the variables in β. We divide the computation of the Hessian into two cases as shown below. Towards this, let ψ_1 be the trigamma function, which is the derivative of the digamma function.

- **Case $c \neq d$:** In this case, using (4.14), we have

$$\frac{\partial}{\partial \beta_{qc}} X_{ip}\mu_{id} = -X_{iq}X_{ip}\mu_{ic}\mu_{id}. \tag{4.18}$$

  Furthermore, since

$$\frac{\partial}{\partial \beta_{qc}}\sum_j \mu_{ij}\psi(\phi_i\mu_{ij}) = \sum_j\left[\psi(\phi_i\mu_{ij})\frac{\partial \mu_{ij}}{\partial \beta_{qc}} + \mu_{ij}\frac{\partial \psi(\phi_i\mu_{ij})}{\partial \beta_{qc}}\right],$$

  using (4.14) and (4.15),

$$\begin{aligned}
\frac{\partial}{\partial \beta_{qc}}\sum_j \mu_{ij}\psi(\phi_i\mu_{ij}) &= X_{iq}\mu_{ic}\cdot\frac{\partial}{\partial \mu_{ic}}\mu_{ic}\psi(\phi_i\mu_{ic}) - X_{iq}\mu_{ic}\sum_j \mu_{ij}\cdot\frac{\partial}{\partial \mu_{ij}}\mu_{ij}\psi(\phi_i\mu_{ij})\\
&= X_{iq}\mu_{ic}\Big(\psi(\phi_i\mu_{ic}) + \phi_i\mu_{ic}\psi_1(\phi_i\mu_{ic}) - \sum_j \mu_{ij}\big(\psi(\phi_i\mu_{ij}) + \phi_i\mu_{ij}\psi_1(\phi_i\mu_{ij})\big)\Big),
\end{aligned}$$

$$\frac{\partial}{\partial \beta_{qc}}\sum_j \mu_{ij}\ln(y_{ij}) = X_{iq}\mu_{ic}\ln(y_{ic}) - X_{iq}\mu_{ic}\sum_j \mu_{ij}\ln(y_{ij}),$$

and

$$\frac{\partial}{\partial\beta_{qc}}\big(\ln(y_{id})-\psi(\phi_i\mu_{id})\big)=-\phi_i X_{iq}\mu_{id}\mu_{ic}\frac{\partial}{\partial\phi_i\mu_{id}}\big(\ln(y_{id})-\psi(\phi_i\mu_{id})\big)=\phi_i X_{iq}\mu_{id}\mu_{ic}\psi_1(\phi_i\mu_{id}).$$

By combining the above, we obtain:

$$\begin{aligned}\frac{\partial}{\partial\beta_{qc}}\frac{\partial}{\partial\beta_{pd}}\ell(y|\mu,\phi)=\sum_{i=1}^{n}\Bigg[&\phi_i X_{ip}\mu_{id}X_{iq}\mu_{ic}\Bigg(\psi(\phi_i\mu_{ic})+\phi_i\mu_{ic}\psi_1(\phi_i\mu_{ic})\\&-\sum_j\mu_{ij}\big(\psi(\phi_i\mu_{ij})+\phi_i\mu_{ij}\psi_1(\phi_i\mu_{ij})\big)\\&-\ln(y_{ic})+\sum_j\mu_{ij}\ln(y_{ij})+\phi_i\mu_{id}\psi_1(\phi_i\mu_{id})\Bigg)\\&-\phi_i X_{ip}\mu_{id}X_{iq}\mu_{ic}\Bigg(\sum_j\mu_{ij}\big(\psi(\phi_i\mu_{ij})-\ln(y_{ij})\big)-\psi(\phi_i\mu_{id})+\ln(y_{id})\Bigg)\Bigg]\\=\sum_{i=1}^{n}\Bigg[&\phi_i X_{ip}\mu_{id}X_{iq}\mu_{ic}\Bigg(\psi(\phi_i\mu_{ic})+\phi_i\mu_{ic}\psi_1(\phi_i\mu_{ic})-\phi_i\sum_j\mu_{ij}^2\psi_1(\phi_i\mu_{ij})\\&-\ln(y_{ic})+2\sum_j\mu_{ij}\ln(y_{ij})+\phi_i\mu_{id}\psi_1(\phi_i\mu_{id})\\&-2\sum_j\mu_{ij}\psi(\phi_i\mu_{ij})+\psi(\phi_i\mu_{id})-\ln(y_{id})\Bigg)\Bigg].\end{aligned}\tag{4.19}$$

- **Case $c=d$:** We now have

$$\frac{\partial}{\partial\beta_{qc}}X_{ip}\mu_{id}=X_{iq}\mu_{ic}(1-\mu_{ic})X_{ip},\tag{4.20}$$

and

$$\frac{\partial}{\partial\beta_{qc}}\big(\ln(y_{ic})-\psi(\phi_i\mu_{ic})\big)=-\phi_i X_{iq}\mu_{ic}(1-\mu_{ic})\psi_1(\phi_i\mu_{ic}).\tag{4.21}$$

Thus,

$$\begin{aligned}\frac{\partial}{\partial\beta_{qc}}\frac{\partial}{\partial\beta_{pc}}\ell(y|\mu,\phi)=\sum_{i=1}^{n}\Bigg[&\phi_i X_{ip}X_{iq}\mu_{ic}^2\Bigg(\psi(\phi_i\mu_{ic})+\phi_i\mu_{ic}\psi_1(\phi_i\mu_{ic})-\sum_j\mu_{ij}\big(\psi(\phi_i\mu_{ij})+\phi_i\mu_{ij}\psi_1(\phi_i\mu_{ij})\big)\\&-\ln(y_{ic})+\sum_j\mu_{ij}\ln(y_{ij})-\phi_i(1-\mu_{ic})\psi_1(\phi_i\mu_{ic})\Bigg)\\&+\phi_i X_{ip}X_{iq}\mu_{ic}(1-\mu_{ic})\Bigg(\sum_j\mu_{ij}\big(\psi(\phi_i\mu_{ij})-\ln(y_{ij})\big)-\psi(\phi_i\mu_{ic})+\ln(y_{ic})\Bigg)\Bigg]\\=\sum_{i=1}^{n}\Bigg[&\phi_i X_{ip}X_{iq}\mu_{ic}^2\Bigg(2\psi(\phi_i\mu_{ic})+2\phi_i\mu_{ic}\psi_1(\phi_i\mu_{ic})-2\sum_j\mu_{ij}\psi(\phi_i\mu_{ij})\\&-\phi_i\sum_j\mu_{ij}^2\psi_1(\phi_i\mu_{ij})-2\ln(y_{ic})+2\sum_j\mu_{ij}\ln(y_{ij})-\phi_i\psi_1(\phi_i\mu_{ic})\Bigg)\\&+\phi_i X_{ip}X_{iq}\mu_{ic}\Bigg(\sum_j\mu_{ij}\big(\psi(\phi_i\mu_{ij})-\ln(y_{ij})\big)-\psi(\phi_i\mu_{ic})+\ln(y_{ic})\Bigg)\Bigg].\end{aligned}\tag{4.22}$$

Furthermore, we have:

$$\frac{\partial^2}{\partial\gamma_k\gamma_m}\ell(y|\mu,\phi)=\sum_{i=1}^{n}\Bigg(\phi_i Z_{ik}\Bigg(\phi_i Z_{im}\psi_1(\phi_i)-\phi_i\sum_j\mu_{ij}Z_{im}\mu_{ij}\psi_1(\phi_i\mu_{ij})\Bigg)$$

$$+ \phi_i Z_{ik} Z_{im} \Big(\psi(\phi_i) + \sum_j \mu_{ij} \big(\ln(y_{ij}) - \psi(\phi_i \mu_{ij}) \big) \Big) \Big)$$
$$= \sum_{i=1}^{n} \phi_i Z_{ik} Z_{im} \Big(\phi_i \psi_1(\phi_i) + \psi(\phi_i) - \phi_i \sum_j \mu_{ij}^2 \psi_1(\phi_i \mu_{ij}) + \sum_j \mu_{ij} \big(\ln(y_{ij}) - \psi(\phi_i \mu_{ij}) \big) \Big). \tag{4.23}$$

Finally, by differentiating 4.7, we can compute the second derivative with respect to β and γ,

$$\begin{aligned}
\frac{\partial^2}{\partial \gamma_k \partial \beta_{pd}} \ell(y|\mu, \phi) &= \sum_{i=1}^{n} Z_{ik} \phi_i \frac{\partial}{\partial \beta_{pd}} \Big(\sum_j \mu_{ij} \big(\ln(y_{ij}) - \psi(\phi_i \mu_{ij}) \big) \Big) \\
&= \sum_{i=1}^{n} Z_{ik} \phi_i \Big(X_{ip} \mu_{id} \frac{\partial}{\partial \mu_{id}} \mu_{id} \big(\ln(y_{id}) - \psi(\phi_i \mu_{id}) \big) \\
&\qquad - \sum_j X_{ip} \mu_{id} \mu_{ij} \frac{\partial}{\partial \mu_{ij}} \mu_{ij} \big(\ln(y_{ij}) - \psi(\phi_i \mu_{ij}) \big) \Big) \\
&= \sum_{i=1}^{n} Z_{ik} \phi_i X_{ip} \mu_{id} \Big(\big(\ln(y_{id}) - \psi(\phi_i \mu_{id}) - \phi_i \mu_{id} \psi_1(\phi_i \mu_{id}) \big) \\
&\qquad - \sum_j \mu_{ij} \big(\ln(y_{ij}) - \psi(\phi_i \mu_{ij}) - \phi_i \mu_{ij} \psi_1(\phi_i \mu_{ij}) \big) \Big).
\end{aligned} \tag{4.24}$$

Appendix 4.B Computation on Dirichlet Distribution with Spatial Lag

Given a matrix of coefficients $\beta \in \mathbb{R}^{K \times J}$, we take

$$E_i^{\Sigma} = \sum_{j'=1}^{J} \exp\Big(\sum_{i'=1}^{n} \sum_{k=K}^{n} M_{ii'}^{-1} X_{i'k} \beta_{kj'} \Big),$$

and define μ as a matrix with elements

$$\mu_{ij} = \frac{\exp\big(\sum_{i'=1}^{n} \sum_{k=1}^{K} M_{ii'}^{-1} X_{i'k} \beta_{kj} \big)}{E_i^{\Sigma}} = \frac{\exp([M^{-1} X \beta]_{ij})}{\sum_{j'=1}^{J} \exp([M^{-1} X \beta]_{ij'})}.$$

In subsections 4.B.1 and 4.B.2, we provide the computations of the first and second derivatives of the loglikelihood function with respect to β, γ and ρ.

4.B.1 First order derivative

Define the matrix $\tilde{X} \in \mathbb{R}^{n \times K}$ as $\tilde{X} = M^{-1} X$. Then the derivative of the log-likelihood with respect to β consists of

$$\frac{\partial}{\partial \beta_{pd}} \ell(y|\mu, \phi) = \sum_{i=1}^{n} \left(\phi_i \tilde{X}_{ip} \mu_{id} \Big(\sum_j \mu_{ij} \big(\psi(\phi_i \mu_{ij}) - \ln(y_{ij}) \big) - \psi(\phi_i \mu_{id}) + \ln(y_{id}) \Big) \right).$$

We further differentiate the log-likelihood with respect to ρ. Note that

$$\begin{aligned}
\frac{\partial M^{-1}}{\partial \rho} &= -M^{-1} \frac{\partial M}{\partial \rho} M^{-1} \\
&= M^{-1} W M^{-1}.
\end{aligned}$$

Hence,

$$\begin{aligned}\frac{\partial}{\partial\rho}(M^{-1}X\beta) &= \frac{\partial M^{-1}}{\partial\rho}X\beta \\ &= M^{-1}WM^{-1}X\beta.\end{aligned}$$

And thus, by defining $U = M^{-1}WM^{-1}X\beta$, we obtain

$$\begin{aligned}\frac{\partial\mu_{ij}}{\partial\rho} &= \frac{U_{ij}\,\exp([M^{-1}X\beta]_{ij})\,\sum_{j'=1}^{J}\exp([M^{-1}X\beta]_{ij'}) - \exp([M^{-1}X\beta]_{ij})\,\sum_{j'=1}^{J}U_{ij'}\exp([M^{-1}X\beta]_{ij'})}{\Big(\sum_{j'=1}^{J}\exp([M^{-1}X\beta]_{ij'})\Big)^2} \\ &= \mu_{ij}U_{ij} - \mu_{ij}\sum_{j'=1}^{J}\mu_{ij'}U_{ij'}. \qquad (4.25)\end{aligned}$$

Using (4.25), we can then compute the derivative of the loglikelihood with respect to ρ for each of the three elements defined in (4.13). Then,

$$\begin{aligned}\frac{\partial A_i}{\partial\rho} &= 0, \\ \frac{\partial B_i}{\partial\rho} &= \sum_j\left(\phi_i\big(\mu_{ij}U_{ij} - \mu_{ij}\sum_{j'=1}^{J}\mu_{ij'}U_{ij'}\big)\frac{\partial}{\partial\phi_i\mu_{ij}}\ln(\Gamma(\phi_i\mu_{ij}))\right) \\ &= \phi_i\sum_j\mu_{ij}U_{ij}\psi(\phi_i\mu_{ij}) - \phi_i\sum_j\mu_{ij}U_{ij}\cdot\sum_j\mu_{ij}\psi(\phi_i\mu_{ij}) \\ &= \phi_i\sum_j\mu_{ij}U_{ij}\Big(\psi(\phi_i\mu_{ij}) - \sum_{j'}\mu_{ij'}\psi(\phi_i\mu_{ij'})\Big), \\ \frac{\partial C_i}{\partial\rho} &= \sum_j\left(\big(\mu_{ij}U_{ij} - \mu_{ij}\sum_{j'=1}^{J}\mu_{ij'}U_{ij'}\big)\frac{\partial}{\partial\mu_{ij}}(\phi_i\mu_{ij}-1)\ln(y_{ij})\right) \\ &= \phi_i\sum_j\mu_{ij}\ln(y_{ij})\Big(U_{ij} - \sum_{j'}\mu_{ij'}U_{ij'}\Big).\end{aligned}$$

By taking the summation of the above, we obtain (4.10).

4.B.2 Second order derivative

Let

$$F_{ij} = \ln(y_{ij})\big(U_{ij} - \sum_{j'}\mu_{ij'}U_{ij'}\big), \qquad (4.26)$$

and

$$G_{ij} = U_{ij}\big(\psi(\phi_i\mu_{ij}) - \sum_{j'}\mu_{ij'}\psi(\phi_i\mu_{ij'})\big). \qquad (4.27)$$

Then,

$$\frac{\partial^2}{\partial\rho^2}\ell(y|\mu,\phi) = \sum_{i=1}^{n}\phi_i\Big(\sum_j\frac{\partial\mu_{ij}}{\partial\rho}\cdot(F_{ij}-G_{ij}) + \sum_j\mu_{ij}\cdot\Big(\frac{\partial}{\partial\rho}F_{ij} - \frac{\partial}{\partial\rho}G_{ij}\Big)\Big). \qquad (4.28)$$

Observe that

$$\begin{aligned}\frac{\partial U}{\partial\rho} &= \frac{\partial}{\partial\rho}(M^{-1}WM^{-1}X\beta) \\ &= \frac{\partial}{\partial\rho}(M^{-1}W)M^{-1}X\beta + M^{-1}W\frac{\partial}{\partial\rho}(M^{-1}X\beta)\end{aligned}$$

$$= 2M^{-1}WM^{-1}WM^{-1}X\beta = 2M^{-1}WU. \tag{4.29}$$

Furthermore, since

$$\begin{aligned}\frac{\partial}{\partial\rho}(\mu_{ij}U_{ij}) &= \mu_{ij}\frac{\partial}{\partial\rho}U_{ij} + U_{ij}\frac{\partial}{\partial\rho}\mu_{ij} \\ &= 2\mu_{ij}V_{ij} + U_{ij}\big(\mu_{ij}U_{ij} - \mu_{ij}\sum_{j'=1}^{J}\mu_{ij'}U_{ij'}\big),\end{aligned} \tag{4.30}$$

by taking $V = M^{-1}WU$, we observe that

$$\begin{aligned}\frac{\partial}{\partial\rho}\sum_j(\mu_{ij}U_{ij}) &= \sum_j\Big(2\mu_{ij}V_{ij} + U_{ij}\big(\mu_{ij}U_{ij} - \mu_{ij}\sum_{j'}\mu_{ij'}U_{ij'}\big)\Big) \\ &= 2\sum_j\big(\mu_{ij}V_{ij}\big) + \sum_j\big(\mu_{ij}U_{ij}^2\big) - \sum_j\big(\mu_{ij}U_{ij}\sum_{j'}\mu_{ij'}U_{ij'}\big) \\ &= \sum_j\mu_{ij}\big(2V_{ij} + U_{ij}^2\big) - \Big(\sum_j\mu_{ij}U_{ij}\Big)^2.\end{aligned} \tag{4.31}$$

Using (4.29) and (4.30), for all i, j,

$$\begin{aligned}\frac{\partial F_{ij}}{\partial\rho} &= \ln(y_{ij})\frac{\partial}{\partial\rho}\big(U_{ij} - \sum_{j'}\mu_{ij'}U_{ij'}\big) \\ &= \ln(y_{ij})\bigg(2V_{ij} - \sum_{j'}\mu_{ij'}\big(2V_{ij'} + U_{ij'}^2\big) + \Big(\sum_{j'}\mu_{ij'}U_{ij'}\Big)^2\bigg),\end{aligned} \tag{4.32}$$

and then using (4.25),

$$\begin{aligned}\frac{\partial G_{ij}}{\partial\rho} &= \frac{\partial U_{ij}}{\partial\rho}\bigg(\psi(\phi_i\mu_{ij}) - \sum_{j'}\mu_{ij'}\psi(\phi_i\mu_{ij'})\bigg) + U_{ij}\,\frac{\partial}{\partial\rho}\bigg(\psi(\phi_i\mu_{ij}) - \sum_{j'}\mu_{ij'}\psi(\phi_i\mu_{ij'})\bigg) \\ &= 2V_{ij}\bigg(\psi(\phi_i\mu_{ij}) - \sum_{j'}\mu_{ij'}\psi(\phi_i\mu_{ij'})\bigg) + U_{ij}\,\Big(\mu_{ij}U_{ij} - \mu_{ij}\sum_{j'=1}^{J}\mu_{ij'}U_{ij'}\Big)\,\frac{\partial}{\partial\mu_{ij}}\psi(\phi_i\mu_{ij}) \\ &\quad - U_{ij}\sum_{j'}\Big(\frac{\partial}{\partial\rho}\mu_{ij'}\psi(\phi_i\mu_{ij'})\Big) \\ &= 2V_{ij}\bigg(\psi(\phi_i\mu_{ij}) - \sum_{j'}\mu_{ij'}\psi(\phi_i\mu_{ij'})\bigg) + U_{ij}\phi_i\,\Big(\mu_{ij}U_{ij} - \mu_{ij}\sum_{j'=1}^{J}\mu_{ij'}U_{ij'}\Big)\,\psi_1(\phi_i\mu_{ij}) \\ &\quad - U_{ij}\sum_{j'}\bigg(\Big(\mu_{ij'}U_{ij'} - \mu_{ij'}\sum_{k=1}^{J}\mu_{ik}U_{ik}\Big)\Big(\psi(\phi_i\mu_{ij'}) + \phi_i\mu_{ij'}\psi_1(\phi_i\mu_{ij'})\Big)\bigg).\end{aligned} \tag{4.33}$$

By introducing $\Omega_{ij} = \mu_{ij}U_{ij}$ and $\Omega_i^{\Sigma} = \sum_j\Omega_{ij}$, and noting $\alpha_{ij} = \phi_i\mu_{ij}$, we combine (4.32) and (4.33) to get

$$\frac{\partial}{\partial\rho}F_{ij} - \frac{\partial}{\partial\rho}G_{ij} = \ln(y_{ij})\Big(2V_{ij} - \sum_{j'}\mu_{ij'}\big(2V_{ij'} + U_{ij'}^2\big) + (\Omega_i^{\Sigma})^2\Big) - 2V_{ij}\Big(\psi(\alpha_{ij}) - \sum_{j'}\mu_{ij'}\psi(\alpha_{ij'})\Big)$$
$$- U_{ij}\phi_i\psi_1(\alpha_{ij})\Big(\Omega_{ij} - \mu_{ij}\Omega_i^{\Sigma}\Big) + U_{ij}\sum_{j'}\Big(\Big(\Omega_{ij} - \mu_{ij}\Omega_i^{\Sigma}\Big)\Big(\psi(\alpha_{ij'}) + \alpha_{ij'}\psi_1(\alpha_{ij'})\Big)\Big).$$

Inserting this expression into (4.28), we can show that $\frac{\partial^2}{\partial\rho^2}\ell(y|\mu,\phi)$ is equal to

$$\sum_{i=1}^{n}\phi_i\Bigg(\sum_j\big(\Omega_{ij} - \mu_{ij}\Omega_i^{\Sigma}\big)(F_{ij} - G_{ij})$$
$$+ \sum_j\mu_{ij}\Bigg(\ln(y_{ij})\Big(2V_{ij} - \sum_{j'}\mu_{ij'}\big(2V_{ij'} + U_{ij'}^2\big) + (\Omega_i^{\Sigma})^2\Big) - 2V_{ij}\Big(\psi(\alpha_{ij}) - \sum_{j'}\mu_{ij'}\psi(\alpha_{ij'})\Big)$$
$$- U_{ij}\phi_i\psi_1(\alpha_{ij})\Big(\Omega_{ij} - \mu_{ij}\Omega_i^{\Sigma}\Big) + U_{ij}\sum_{j'}\Big(\Big(\Omega_{ij} - \mu_{ij}\Omega_i^{\Sigma}\Big)\Big(\psi(\alpha_{ij'}) + \alpha_{ij'}\psi_1(\alpha_{ij'})\Big)\Big)\Bigg)\Bigg).$$

Furthermore, differentiating (4.10) with respect to γ_k gives

$$\frac{\partial^2}{\partial\rho\partial\gamma_k}\ell(y|\mu,\phi) = \sum_{i=1}^{n}\Bigg(Z_{ik}\phi_i\sum_j\mu_{ij}\Big(\ln(y_{ij})\Big(U_{ij} - \sum_{j'}\mu_{ij'}U_{ij'}\Big) - U_{ij}\Big(\psi(\phi_i\mu_{ij}) - \sum_{j'}\mu_{ij'}\psi(\phi_i\mu_{ij'})\Big)\Big)$$
$$- \phi_i\sum_j\mu_{ij}U_{ij}\Big(Z_{ik}\phi_i\mu_{ij}\psi_1(\phi_i\mu_{ij}) - \sum_{j'}Z_{ik}\phi_i\mu_{ij'}^2\psi_1(\phi_i\mu_{ij'})\Big)\Bigg)$$
$$= \sum_{i=1}^{n}Z_{ik}\phi_i\Bigg(\sum_j\mu_{ij}\Big(\ln(y_{ij})\Big(U_{ij} - \sum_{j'}\mu_{ij'}U_{ij'}\Big) - U_{ij}\Big(\psi(\phi_i\mu_{ij}) - \sum_{j'}\mu_{ij'}\psi(\phi_i\mu_{ij'})\Big)$$
$$- \phi_iU_{ij}\Big(\mu_{ij}\psi_1(\phi_i\mu_{ij}) - \sum_{j'}\mu_{ij'}^2\psi_1(\phi_i\mu_{ij'})\Big)\Big)$$
$$= \sum_{i=1}^{n}Z_{ik}\phi_i\Bigg(\sum_j\mu_{ij}\Big(\ln(y_{ij})\Big(U_{ij} - \Omega_i^{\Sigma}\Big) - U_{ij}\Big(\psi(\alpha_{ij}) - \sum_{j'}\mu_{ij'}\psi(\alpha_{ij'})\Big)$$
$$- \phi_iU_{ij}\Big(\mu_{ij}\psi_1(\alpha_{ij}) - \sum_{j'}\mu_{ij'}^2\psi_1(\alpha_{ij'})\Big)\Big)\Bigg).$$

Finally, to compute the derivative of (4.10) with respect to β_{pd}, we need to compute the derivatives of U_{ij} and $\sum_j\mu_{ij}U_{ij}$. First, we define $Q = M^{-1}WM^{-1}X$. Thus, $U = Q\beta$, that is, $U_{ij} = \sum_k Q_{ik}\beta_{kj}$. Hence, $\frac{\partial U_{ij}}{\partial\beta_{pd}} = Q_{ip}$ if $d = j$, and 0 otherwise. Hence, by adapting (4.14) and (4.15) to $\tilde{X}$, we have

$$\frac{\partial}{\partial\beta_{pd}}\mu_{ij}U_{ij} = \begin{cases} Q_{ip}\mu_{id} + U_{id}\tilde{X}_{ip}\mu_{id}(1-\mu_{id}), \text{ if } d = j. \\ -U_{ij}\tilde{X}_{ip}\mu_{ij}\mu_{id}, \text{ if } d \neq j. \end{cases} \tag{4.34}$$

$$\frac{\partial}{\partial\beta_{pd}}\sum_j\mu_{ij}U_{ij} = \mu_{id}\Big(Q_{ip} + \tilde{X}_{ip}\big(U_{id} - \sum_j\mu_{ij}U_{ij}\big)\Big)$$
$$= \mu_{id}\Big(Q_{ip} + \tilde{X}_{ip}\big(U_{id} - \Omega_i^{\Sigma}\big)\Big). \tag{4.35}$$

In order to be more concise, we keep the same notations of F_{ij} and G_{ij} defined in (4.26) and (4.27).

$$\frac{\partial^2}{\partial\rho\partial\beta_{pd}}\ell(y|\mu,\phi) = \sum_{i=1}^{n}\phi_i\Big(\sum_j\frac{\partial\mu_{ij}}{\partial\beta_{pd}}\cdot(F_{ij} - G_{ij}) + \sum_j\mu_{ij}\cdot\Big(\frac{\partial}{\partial\beta_{pd}}F_{ij} - \frac{\partial}{\partial\beta_{pd}}G_{ij}\Big)\Big)$$

$$= \sum_{i=1}^{n} \phi_i \Big(\tilde{X}_{ip} \mu_{id} \Big((F_{id} - G_{id}) - \sum_j \mu_{ij} (F_{ij} - G_{ij}) \Big) + \sum_j \mu_{ij} \frac{\partial}{\partial \beta_{pd}} F_{ij} - \sum_j \mu_{ij} \frac{\partial}{\partial \beta_{pd}} G_{ij} \Big). \tag{4.36}$$

Then,

$$\sum_j \mu_{ij} \frac{\partial}{\partial \beta_{pd}} F_{ij} = \mu_{id} \ln(y_{id}) Q_{ip} - \mu_{id} \Big(Q_{ip} + \tilde{X}_{ip} \big(U_{id} - \Omega_i^{\Sigma} \big) \Big) \sum_j \mu_{ij} \ln(y_{ij}),$$

and

$$\begin{aligned}
\sum_j \mu_{ij} \frac{\partial}{\partial \beta_{pd}} G_{ij} &= \sum_j \mu_{ij} \Big(\psi(\phi_i \mu_{ij}) - \sum_{j'} \mu_{ij'} \psi(\phi_i \mu_{ij'}) \Big) \frac{\partial}{\partial \beta_{pd}} U_{ij} \\
&\quad + \sum_j \mu_{ij} U_{ij} \frac{\partial}{\partial \beta_{pd}} \big(\psi(\phi_i \mu_{ij}) - \sum_{j'} \mu_{ij'} \psi(\phi_i \mu_{ij'}) \big) \\
&= Q_{ip} \mu_{id} \Big(\psi(\phi_i \mu_{id}) - \sum_{j'} \mu_{ij'} \psi(\phi_i \mu_{ij'}) \Big) \\
&\quad + \sum_j \mu_{ij} U_{ij} \frac{\partial}{\partial \beta_{pd}} \psi(\phi_i \mu_{ij}) - \sum_j \mu_{ij} U_{ij} \sum_j \frac{\partial}{\partial \beta_{pd}} \mu_{ij} \psi(\phi_i \mu_{ij}) \\
&= Q_{ip} \mu_{id} \Big(\psi(\phi_i \mu_{id}) - \sum_j \mu_{ij} \psi(\phi_i \mu_{ij}) \Big) \\
&\quad + \tilde{X}_{ip} \mu_{id} \phi_i \Big(\mu_{id} U_{id} \psi_1(\phi_i \mu_{id}) - \sum_j \mu_{ij}^2 U_{ij} \psi_1(\phi_i \mu_{ij}) \Big) \\
&\quad - \sum_j (\mu_{ij} U_{ij}) \cdot \tilde{X}_{ip} \mu_{id} \Big(\psi(\phi_i \mu_{id}) + \phi_i \mu_{id} \psi_1(\phi_i \mu_{id}) - \sum_j \mu_{ij} \big(\psi(\phi_i \mu_{ij}) + \phi_i \mu_{ij} \psi_1(\phi_i \mu_{ij}) \big) \Big) \\
&= \mu_{id} \Bigg(Q_{ip} \Big(\psi(\alpha_{id}) - \sum_j \mu_{ij} \psi(\alpha_{ij}) \Big) + \tilde{X}_{ip} \phi_i \Big(\Omega_{id} \psi_1(\alpha_{id}) - \sum_j \mu_{ij}^2 U_{ij} \psi_1(\alpha_{ij}) \Big) \\
&\quad - \tilde{X}_{ip} \Omega_i^{\Sigma} \Big(\psi(\alpha_{id}) + \alpha_{id} \psi_1(\alpha_{id}) - \sum_j \mu_{ij} \big(\psi(\alpha_{ij}) + \alpha_{ij} \psi_1(\alpha_{ij}) \big) \Big) \Bigg).
\end{aligned}$$

Finally,

$$\begin{aligned}
\frac{\partial^2}{\partial \rho \partial \beta_{pd}} \ell(y|\mu, \phi) = \sum_{i=1}^{n} \phi_i \mu_{id} \Bigg(& \tilde{X}_{ip} \Big((F_{id} - G_{id}) - \sum_j \mu_{ij} (F_{ij} - G_{ij}) \Big) \\
& + \ln(y_{id}) Q_{ip} - \Big(Q_{ip} + \tilde{X}_{ip} \big(U_{id} - \Omega_i^{\Sigma} \big) \Big) \sum_j \mu_{ij} \ln(y_{ij}) \\
& - Q_{ip} \Big(\psi(\alpha_{id}) - \sum_j \mu_{ij} \psi(\alpha_{ij}) \Big) - \tilde{X}_{ip} \phi_i \Big(\Omega_{id} \psi_1(\alpha_{id}) - \sum_j \mu_{ij}^2 U_{ij} \psi_1(\alpha_{ij}) \Big) \\
& + \tilde{X}_{ip} \Omega_i^{\Sigma} \Big(\psi(\alpha_{id}) + \alpha_{id} \psi_1(\alpha_{id}) - \sum_j \mu_{ij} \big(\psi(\alpha_{ij}) + \alpha_{ij} \psi_1(\alpha_{ij}) \big) \Big) \Bigg).
\end{aligned} \tag{4.37}$$

Appendix 4.C Equivalence between crossentropy and multinomial

Let $J \in \mathbb{N}$ be the number of classes and $K \in \mathbb{N}$ the number of features. Let $Y_i = (Y_{i1}, \ldots, Y_{iJ}) \sim \text{Mult}(n_i, p_{i1}, \ldots, p_{iJ})$, where $\sum_j p_{ij} = 1$ and $\sum_j Y_{ij} = n_i$ with Y_{ij} non

negative integers. The probability mass function is

$$P[Y_{i1} = y_{i1}, \dots, Y_{iJ} = y_{iJ}] = c(n_i) \prod_{j=1}^{J} p_{ij}^{y_{ij}}$$

where $c(n_i)$ is a term depending on n_i.

Now consider the multinomial logit model where we link the p_j to some covariate X. Let's define X_i the K-dimensional vector for the sample i. Then, for $\beta_j \in \mathbb{R}^K$,

$$p_{ij} = \frac{\exp(X_i^T \beta_j)}{\sum_{j'} \exp(X_i^T \beta_{j'})}. \tag{4.38}$$

Let $\theta = (\beta_1, \dots, \beta_J)$. The loglikelihood for a sample of N observations is

$$\begin{aligned} \ell(\theta; y_1, \dots, y_n) &= \sum_{i=1}^{N} \ln(c(n_i)) + \sum_{i=1}^{N} \sum_{j=1}^{J} y_{ij} \ln(p_{ij}) \\ &\propto \sum_{i=1}^{N} \sum_{j=1}^{J} y_{ij} \ln(p_{ij}), \end{aligned}$$

and, by noting $\tilde{y_{ij}} = \frac{y_{ij}}{n_i} \in [0, 1]$,

$$\begin{aligned} \ell(\theta; y_1, \dots, y_n) &\propto \sum_{i=1}^{N} \sum_{j=1}^{J} n_i \tilde{y_{ij}} \ln(p_{ij}) \\ &\propto \sum_{i=1}^{N} n_i \sum_{j=1}^{J} \tilde{y_{ij}} \ln(p_{ij}). \end{aligned}$$

Then, in the case where all the samples have the same size m, we have $n_i = m$ for all i. From this we have that

$$\begin{aligned} \ell(\theta; y_1, \dots, y_n) &\propto m \sum_{i=1}^{N} \sum_{j=1}^{J} \tilde{y_{ij}} \ln(p_{ij}) \\ &\propto \sum_{i=1}^{N} \sum_{j=1}^{J} \tilde{y_{ij}} \ln(p_{ij}) = -CE(\tilde{y}, p). \end{aligned}$$

Appendix 4.D Results of the multinomial model on the synthetic dataset

Tables 4.9, 4.10 and 4.11 present the bias, standard deviation and mean squared error of the estimator using the cross-entropy minimization on the synthetic Dirichlet generated data described in Section 4.3.1. Table 4.12 describes the mean performances of the evaluated parameters (with $n = 1000$) on a test set.

Table 4.9: Bias, standard deviation and mean squared error of the estimated parameters with the cross-entropy minimization ($\rho = 0.1$). The results are presented as the mean of the differences on the 100 iterations, the standard deviation within parenthesis, and the mean squared error within square brackets.

	Spatial			Not spatial		
Parameter	n=50	n=200	n=1000	n=50	n=200	n=1000
β_{01}	0.019 (0.088) [0.008]	0.002 (0.051) [0.003]	0.002 (0.022) [0.0]	0.02 (0.102) [0.011]	0.0 (0.058) [0.003]	0.002 (0.025) [0.001]
β_{02}	0.021 (0.108) [0.012]	0.004 (0.048) [0.002]	0.003 (0.022) [0.0]	0.037 (0.13) [0.018]	0.022 (0.056) [0.004]	0.016 (0.027) [0.001]
β_{11}	-0.052 (0.117) [0.016]	-0.02 (0.051) [0.003]	-0.002 (0.023) [0.001]	-0.052 (0.117) [0.016]	-0.018 (0.051) [0.003]	-0.001 (0.024) [0.001]
β_{12}	0.141 (0.127) [0.036]	0.043 (0.076) [0.008]	0.008 (0.035) [0.001]	0.145 (0.129) [0.038]	0.046 (0.077) [0.008]	0.013 (0.036) [0.001]
β_{21}	0.016 (0.12) [0.015]	0.008 (0.051) [0.003]	-0.001 (0.02) [0.0]	0.016 (0.118) [0.014]	0.006 (0.05) [0.003]	-0.002 (0.02) [0.0]
β_{22}	0.091 (0.126) [0.024]	0.031 (0.063) [0.005]	0.006 (0.031) [0.001]	0.091 (0.128) [0.025]	0.031 (0.065) [0.005]	0.01 (0.033) [0.001]
ρ	-0.006 (0.068) [0.005]	-0.003 (0.033) [0.001]	0.001 (0.013) [0.0]	/ /	/ /	/ /

Table 4.10: Bias, standard deviation and mean squared error of the estimated parameters with the cross-entropy minimization ($\rho = 0.5$). The results are presented as the mean of the differences on the 100 iterations, the standard deviation within parenthesis, and the mean squared error within square brackets.

	Spatial			Not spatial		
Parameter	n=50	n=200	n=1000	n=50	n=200	n=1000
β_{01}	-0.001 (0.058) [0.003]	0.006 (0.031) [0.001]	0.001 (0.013) [0.0]	0.013 (0.216) [0.047]	0.015 (0.115) [0.013]	0.022 (0.051) [0.003]
β_{02}	-0.005 (0.062) [0.004]	0.007 (0.027) [0.001]	0.001 (0.013) [0.0]	0.123 (0.377) [0.157]	0.206 (0.195) [0.08]	0.169 (0.086) [0.036]
β_{11}	-0.036 (0.104) [0.012]	-0.011 (0.052) [0.003]	-0.002 (0.026) [0.001]	-0.026 (0.139) [0.02]	0.001 (0.069) [0.005]	0.007 (0.033) [0.001]
β_{12}	0.194 (0.121) [0.052]	0.055 (0.082) [0.01]	0.015 (0.036) [0.002]	0.314 (0.211) [0.143]	0.211 (0.112) [0.057]	0.204 (0.054) [0.045]
β_{21}	0.033 (0.081) [0.008]	0.009 (0.05) [0.003]	0.004 (0.023) [0.001]	0.017 (0.117) [0.014]	-0.021 (0.066) [0.005]	-0.023 (0.031) [0.001]
β_{22}	0.121 (0.114) [0.028]	0.036 (0.066) [0.006]	0.01 (0.03) [0.001]	0.174 (0.194) [0.068]	0.103 (0.102) [0.021]	0.11 (0.048) [0.014]
ρ	-0.008 (0.036) [0.001]	0.0 (0.015) [0.0]	0.001 (0.007) [0.0]	/ /	/ /	/ /

Table 4.11: Bias, standard deviation and mean squared error of the estimated parameters with the cross-entropy minimization ($\rho = 0.9$). The results are presented as the mean of the differences on the 100 iterations, the standard deviation within parenthesis, and the mean squared error within square brackets.

	Spatial			Not spatial		
Parameter	n=50	n=200	n=1000	n=50	n=200	n=1000
β_{01}	0.002 (0.063) [0.004]	0.0 (0.009) [0.0]	0.0 (0.004) [0.0]	0.143 (1.442) [2.099]	0.062 (0.73) [0.537]	0.179 (0.292) [0.117]
β_{02}	-0.004 (0.088) [0.008]	-0.003 (0.01) [0.0]	-0.0 (0.004) [0.0]	0.696 (1.836) [3.853]	0.845 (0.743) [1.267]	0.726 (0.34) [0.643]
β_{11}	-0.188 (0.316) [0.135]	-0.029 (0.054) [0.004]	-0.003 (0.023) [0.001]	-0.22 (0.37) [0.185]	-0.179 (0.155) [0.056]	-0.218 (0.067) [0.052]
β_{12}	0.643 (0.355) [0.54]	0.164 (0.112) [0.039]	0.033 (0.038) [0.003]	1.109 (0.426) [1.411]	1.291 (0.199) [1.705]	1.32 (0.099) [1.752]
β_{21}	0.137 (0.229) [0.071]	0.031 (0.052) [0.004]	0.005 (0.021) [0.0]	0.103 (0.305) [0.104]	0.065 (0.144) [0.025]	0.079 (0.058) [0.01]
β_{22}	0.5 (0.328) [0.358]	0.118 (0.081) [0.021]	0.025 (0.029) [0.001]	0.76 (0.409) [0.745]	0.802 (0.162) [0.669]	0.855 (0.083) [0.739]
ρ	-0.004 (0.02) [0.0]	0.0 (0.003) [0.0]	-0.0 (0.001) [0.0]	/ /	/ /	/ /

Table 4.12: Scores between the μ^* of the test set, and the estimated $\hat{\mu}$ (computed with the parameters estimated with $n = 1000$ with the minimization of cross-entropy). The results are displayed as mean on the 100 iterations (standard deviation).

	Model	$\boldsymbol{R^2}$	**RMSE**	**Cross-entropy**	**Cos similarity**
$\rho = 0.1$	Not spatial	0.9314 ($< 10^{-4}$)	0.0736 ($< 10^{-4}$)	0.6655 (0.0001)	0.9856 ($< 10^{-4}$)
	Spatial	0.9338 ($< 10^{-4}$)	0.072 ($< 10^{-4}$)	0.6644 (0.0001)	0.9862 ($< 10^{-4}$)
$\rho = 0.5$	Not spatial	0.8421 (0.0001)	0.1207 ($< 10^{-4}$)	0.6764 (0.0002)	0.9618 ($< 10^{-4}$)
	Spatial	0.9414 ($< 10^{-4}$)	0.07 ($< 10^{-4}$)	0.627 (0.0002)	0.9872 ($< 10^{-4}$)
$\rho = 0.9$	Not spatial	0.274 (0.0021)	0.3121 (0.0002)	0.8576 (0.0019)	0.7877 (0.0003)
	Spatial	0.9761 ($< 10^{-4}$)	0.0535 ($< 10^{-4}$)	0.3716 (0.0009)	0.9933 ($< 10^{-4}$)

Conclusion

In this chapter, we introduced a new model, the Spatial AutoRegressive (SAR) model for a Dirichlet distribution, as a solution to address spatial characteristics in compositional data. This model was compared to its non-spatial counterpart and exhibited better performances in both synthetic dataset and real life datasets.

Additionally, we compared the SAR model to a spatial multinomial model for compositional data, which yielded superior performance in practical applications. We hypothesize that this effect comes from the unique nature of our datasets, where the compositionality is recreated. For instance, in the Maupiti dataset, the labels represent the cumulative count of pixels in each class, rather than a natural proportion. This peculiarity could elucidate the SAR Dirichlet's comparatively lower performance compared to the spatial multinomial for compositional data.

In light of these results, although the SAR Dirichlet outperforms the non-spatial model, it falls short of the performance thresholds required for its inclusion in our final framework. Consequently, it will not be incorporated into the tool, as we prioritize other models demonstrating superior performance.

CHAPTER 5

Automated satellite mapping of seabed classification for coral reef-lagoon systems

The work in this chapter is in preparation for a submission to Coral Reefs: T. Nguyen, D. Sous, K. Mengersen, B. Liquet, S. Meulé and F. Bouchette. Automated satellite mapping of structure-based seabed classification for coral reef-lagoon systems, application to the Maupiti island.

Synopsis

The objective of this chapter is to present the final satellite-based coral reef mapping tool that was developed during this thesis.

The workflow consists in combining a pixel-based and an object-based method to increase the accuracy of the models. These models are a Random Forest classifier and regressor, trained with the Maupiti dataset. Furthermore, to take into account the spatial effects without modifying the models, the spatial information of the neighbours are added as supplementary covariates when training the model.

Two distinct models are trained, each tailored to accommodate diverse satellite sources: Pleiades and Sentinel-2 satellites. The proposed models can then predict the mapping for a given input satellite image.

The workflow is adaptable and can be retrained by any user who has suitable training data.

5.1 Introduction

Coral reefs and related lagoon systems are now experiencing a planetary degradation [27], [183], [356]. An increasing research effort is engaged to track their evolution in time and space, in particular with large-extent remote mapping techniques such as satellite-based surveys [37], [63], [298]. A global challenge of coral reef mapping is to identify and to delineate the different seabed zones, allowing to monitor their response to climatic evolution or human actions [16], [182], [442]. Each mapping study requires initially to define a classification of seabed types, i.e. to build a list of distinct classes in which each portion of the real-world seabed will be assigned to.

The class definition historically relies on a series of criteria built on biologic, ecological, historical, geomorphic and geological issues, depending on the available data and the scope and scales of each study [216]. The main approaches, generally referred to as geomorphic or geomorphological classification, rely on the definition of distinct seascape types in reef lagoon system, e.g. reef slope, reef crest or lagoon. Based on such types of classification, several mapping at regional and/or global scales have been developed [13], [31], [188], [302], [338], [400]. The matching of class definition and nomenclature is not perfect, and the presence of different levels in most classification increases the difficulty to reach a common practice. The efforts spent by [216], in the Reef Cover framework, to unify existing approaches must be emphasized. Considering the cost of retrieving detailed field observations, in particular when dealing with vast geographic areas [216], [241] and/or survey at high temporal frequency, automated mapping tools can be developed to overcome the limitations of man-made zonation. Satellite-based mappings are overwhelmingly used owing to their unique ability to cover wide spatial extents at relevant resolution [298]. Built on regional/global geomorphic classifications, several satellite-based international database projects are useful illustrations of the satellite tool capabilities, see *Allen Coral Atlas* https://allencoralatlas.org/atlas/#12.10/-16.4507/-152.2511, the *Khaled bin Sultan Living Oceans Foundation*, and the *Millennium Coral Reef Mapping Project*, http://www.imars.usf.edu/millennium-coral[13]. A first major issue when deploying a satellite mapping project is the choice of the satellite data source, based on a compromise between spatial and spectral resolutions, image cost and space and time covering [298]. Additionally, satellite imagery has inherent limitations, including issues like sunglint [176], atmospheric light scattering [156], [173], and the common use of preprocessing techniques such as geometric [140] and radiometric [313], [315] corrections before it becomes usable. Moreover, satellite images are vulnerable to cloud coverage, rendering some images inapplicable [118], [211]. Once the image selection is meticulously achieved, and often associated to study-specific preprocessing steps, the use of machine learning is particularly well-suited for automating labor-intensive tasks like seabed classification [63], [114], [416]. Two distinct methods are commonly employed for this mapping [8], [20]: pixel-based classification, which classifies individual pixels, and object-based classification, which classifies groups of pixels. The former, even though easier to implement, generally provides lower performances than the latter [8], [20], [128], [460], because the object-based classification allows to take into account the information contained in the neighbours. The machine learning models most commonly used are the Support Vector Machine [162], [329], [438], Decision Trees [329], [438], Neural Networks [82] and Random Forest [42], [329], [438], the latter being usually associated to higher performance [298].

While often combined with the geomorphic approach, a second type of classification relies on the characterization of benthic habitat. The underlying motivation is that the architectural complexity of the seabed is a major controlling factor not only for the benthic habitat, because being related to the health and the richness of the reef ecosystem [64], [161], [326] or even being used as a restoration tool [451], but also for the hydrodynamical functioning of the reef-lagoon system, the roughness being a governing parameter for wave sheltering, water level, circulation and residence time [184], [256], [393], [395], [399], [424] and for the investigation of the founding principles of historical reef formation [334]–[336], [368], [394]. Benthic classifications generally need high-resolution (sub-metric) data inferred from field survey, and are therefore limited in terms of spatial covering [171]. In the next years, the continuous technological advancements may lead to the merging of scaling, i.e. downscaling/upscaling of geomorphic/benthic approaches [151], [216].

The aim of the present study is twofold. First, we aim to define a classification mainly based on topography-based features of the seabed. This approach can be considered as a benthic classification, but we wish to emphasize that the focus is made on the architectural structure of seabed, i.e. potentially related to physical metrics of the bed geometry [394], rather than directly connected to habitat-related ecological issues. This expert-based classification relies on a series of field surveys, using both qualitative observations and quantitative measurements. The analysis is carried out on a single study site, the Maupiti island in French Polynesia, selected for the richness of its seabed structure [394]. The second aim of the present study is to present a high performance satellite mapping tool, in order to assess the extent to which the topography-based classification inferred from field surveys can be remotely inferred from automated satellite mapping and therefore to help filling the gap between geomorphic and benthic mapping approaches. The proposed tool combines pixel- and object-based approaches [199], [458] in an original way. The first part of the paper is dedicated to the description of the expert-based zonation and the developed workflow, in Sections 5.2.2 and 5.2.3, respectively. The results are presented in Section 5.3 and discussed in Section 5.4.

5.2 Materials and methods

5.2.1 Study site

Maupiti ("the Stucked Twins") is a diamond-shaped island located in the western part of the Society archipelago in French Polynesia. It displays a high volcanic central island bordered by two barrier reefs at east and south-west sides, two emerged vegetated areas ("motus") on the north side separated by a system of shallow breaches and a narrow but deep pass at the southern end.

The Maupiti island has been selected as study site owing to the facts that (i) it encompasses a wide variety of seabed types, from pure sand areas to well-developed reticulate reef systems, (ii) it can be considered as a nearly untouched coral reef-lagoon island, with very few engineered areas and (iii) it has been monitored by a series of recent field campaign during the MAUPITI HOE project (2018-2022) [394], [397], [398].

5.2.2 Expert-based zonation

Strategy

The expert-based zonation performed on the Maupiti system was built on three complementary tools. It was first based on the combination of underwater observations consisting in a series of underwater pictures, videos and survey notes taken in 2018 and 2021. The second data source is a Pleiades satellite 4-band image captured on June, 14 2021. Third, a series of high resolution bathymetric survey have been performed on the field. For this data part, most of the efforts have been placed on the documentation on the SW barrier reef. The observations highlighted the fractal nature of the living coral reef colonies and the variability of the reef elevation statistics, quantified by high-order statistical moments such as the skewness [394]. The final complete mapping of the island is displayed in Figure 5.6.

Classes definition

We defined first two meta-classes, namely Sand and Coral, based on the dominant type observed on a given area. Each meta-class is then subdivided in a series of classes described hereafter. The actual class name concatenates the first letter of the meta-class (S or C) with three letters for the class itself. The Sand metaclass is equivalent to the area called "reef aprons" defined by [343] for maupiti island, which include bare sand to gravel with scattered rocky patches, coral heads, and small patch reefs. The related classes are from a single dominant sand [S-S] area to coral/sand mixtures [S-LSC]. The Coral metaclass gather classes dominated by coral. An additional Deepwater meta-class is used to describe areas where depth is too large to allow satellite-based mapping.

- The **Spur And Grooves [C-SAG]** class refers to a common and striking feature of shallow fore-reef areas worldwide: *"Linear ridges and channels, which are usually oriented perpendicular to the reef crest and/or the incoming waves, form a comb-tooth pattern that may develop a relief of 6–8 m and reach from the reef crest down to water depths of 20 m"* [143] (Fig. 5.1H). Research efforts are engaged to better understand their formation and history [108], [143], [382] and their effect of wave transformation and water circulation [107], [354], [355], [384].

- The **Full Coral class [C-FCO]** refers to the very compact, stubby, small scale, living coral bed in very shallow or partly emerged wave-exposed reef-crest environments (Fig. 5.1A). This is the higher and shallower part of the coral-algal zone (algal ridge/crest) [232]. The reef geometrical structure is quite regular, with a typical dimensional scale about 20 cm both in height and length of reef elements and spacing [394]. This nearly isotropic structure of the FCO class is associated to a weak skewness of the bed elevation: highs and troughs of the reef colony are rather homogeneously distributed around the mean bed elevation. The extension of the FCO class generally coincides with the reach of breaking waves.

- Sand patches are observed in the **Mixed Coral Dominance class [C-MCD]**, but the coral cover remains dominant. C-MCD is mainly found on the reef flat. This is the first zone of the transitional area from the C-FCO class over the reef crest to sandy

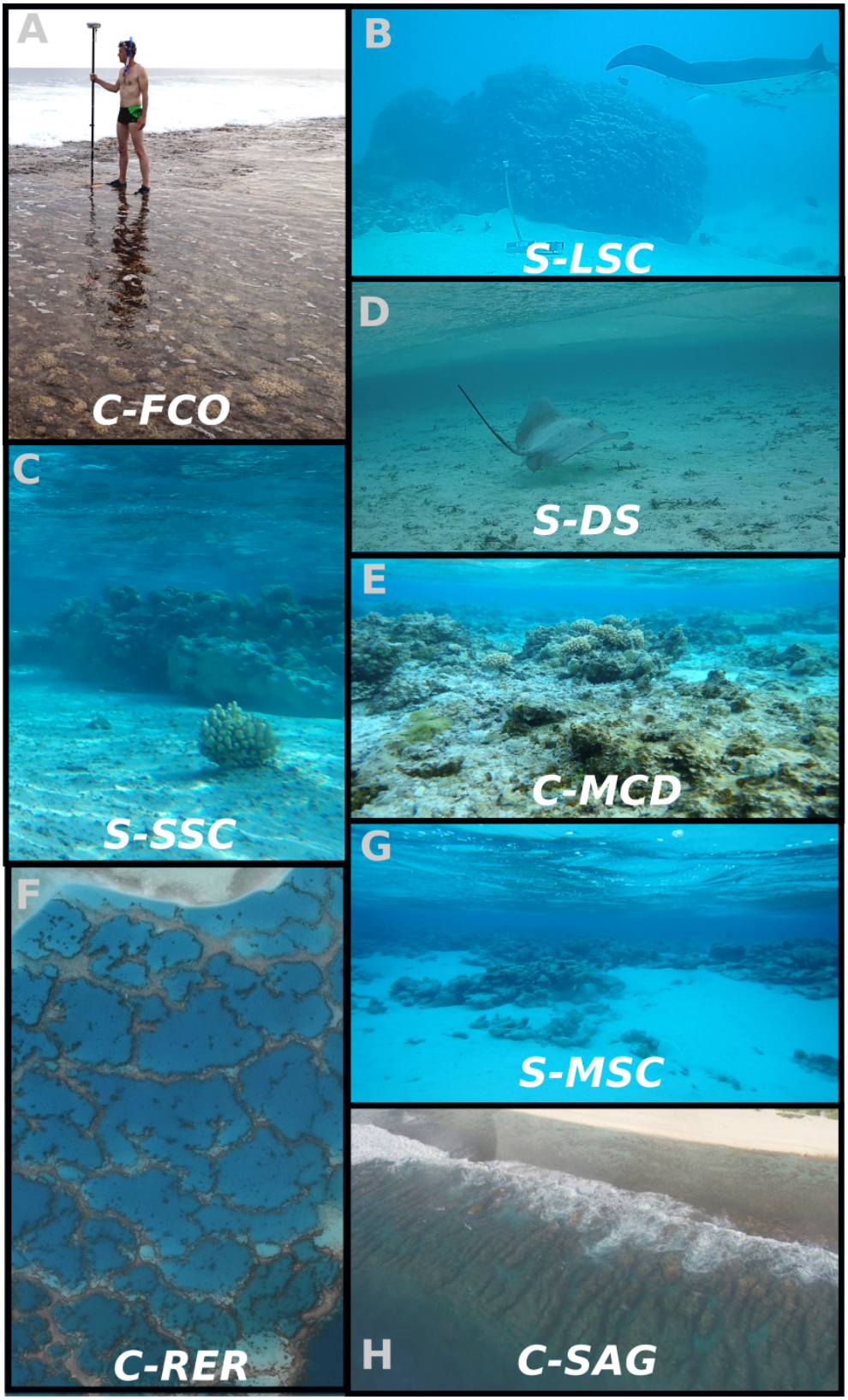


Figure 5.1: Illustrative pictures of several seabed classes. A: C-FCO. B: S-LSC. C: S-SSC. D: S-DS. E: C-MCD. F: C-RER. G: S-MSC. H: C-SAG

areas of the lagoon. The C-MCD displays an hummocky surface but, by contrast to the C-FCO class, the bed coverage by coral reef is not complete, the colony being cut by different gaps and narrows and covered by a film of sand [232] (Fig. 5.1E). The height scale of the reef colony typically ranges between 0.4 to 0.8 m. Typical depth for [C-MCD] class is between 0.5 m to 2 m.

- The **Reticulate reef - [C-RER]** class, also called cellular [50], are defined as a network of coral-sediment ridges resulting from the "*mixture of in situ growth and detrital sediment on coral-rimmed reticulate ridges. Deeper, narrow ridges consist of coral-algal structures only; most shallow, broader ridges consist of a double row of coralgal structures enclosing loose sediment*" [368]. Typical depth for C-RER class in Maupiti is between 5 m to 10 m, the ridge top being nearly at the water surface. Two distinct length scales should be used to characterise the reticulate reefs, one for the rim cell size, typically between 50 to 200m, and a much smaller one for the ridge, typically between 3 and 10m (Figure 5.1F). The C-RER structure is therefore generally marked by a strongly asymmetric (positively skewed) distribution of bed elevation.

- The **Shore Reef - [C-SHR]** class describes the shallow fringing reef as defined by [343], mostly flanked to the main island. It is a compact reef, with decimetric bed fluctuations, generally living in high turbidity waters in 0 to 2 m depth.

- The **Transitional Shore Reef - [C-TSR]** class is a transitional structure between C-RER and C-SHR. Typical depth for [C-TSR] class is between 0 m to 5 m.

- The **Abraded Emerged Reef class [C-AER]** class is a zone of beach rock mostly rising above sea level along the seaward side of the motu. The exposed parts of this abraded reef can be episodically exposed to wave action, in particular during high wave / high water level periods. This class may be associated with raised beach rocks, emerged marine notches or abraded reef flats depending on their elevation above sea level at the time of their formation [67], [323].

- The **Coral Dredging - [C-CDR]** class corresponds to man-engineered area related to infrastructure development, leading to a massive destruction/restructuration of the seabed. In Maupiti, we can observed coral dredging on area composed by Reticulate reef, Transitional Shore Reef and Shore Reef near the harbour and the old pearl oyster farm. Typical depth for [C-CDR] class is between 0 m to 2 m.

- The **Sand [S-SS] class** corresponds to shallow water, nearly pure sand bed. The bed geometry combines here the sand grain size (typically lower than 1mm) and wave- or current-generated sand ripples with a typical height of a few centimetres.

- The **Depleted Sand [S-DS]** class encompasses shallow water sand bed areas partly abraded or depleted, leaving the dead reef substratum partly apparent and sparsely covered by coral debris (Fig. 5.1D).

- The **Mixed Sand Coral [S-MSC]** class is structurally close to the C-MCD, i.e. dispatched coral reef elements separated by sand-covered lows and throughs and typical depths are similar (0.5-2.5m). The main difference being that the reef coverage for S-MSC is lower, typically between 5 and 50 % (Fig. 5.1G). This results in a higher

positive skewness of the bed elevation [394]. The typical height of the reef pinnacles is of the order of 1m.

- Intermediate class between S-MSC and S-S or S-DS, the **Shallow Sparse Coral [S-SSC]** class is characterised by sandy seabed covered with sparse meter-scaled coral pinnacles (Fig. 5.1C). The reef cover is lower than 5 %. The bed elevation distribution is even more positively skewed compared to the S-MSC class.

- The **Large Sparse Coral [S-LSC]** class describes intermediate to deep water sand bed area dotted with sparse meter to plurimeter-scaled coral pinnacles with very shallow colony top (Fig. 5.1B). Typical depth for [S-LSC] class is between 5 m to 15 m.

- In the **Deep Water metaclass [DW]** class, water depth greater than 15 m. No field data has been collected in the area, and satellite images cannot give any information about this deep water.

5.2.3 Automated satellite detection tool

Strategy overview

A short introductory overview of the method is given here while technical details are provided in the following sections. Note first that we designed our tool to be able to deal with a raw satellite image, without any preprocessing technique which generally requires a ground-truth knowledge. The overall aim is to develop a tool able to provide a complete mapping of the study site, i.e. to predict the spatial distribution of classes. Figure 5.2 displays a schematic operation diagram of the mapping process. The final product of the tool is to assign a *label* at each *segment* of the map. The segments are subdivisions of the initial image provided by a decomposition, i.e. the *segmentation*, of the image into representative geomorphological areas (Section 5.2.3). The labels provided by the mapping tool contain the expected proportion of each class for the considered segment, i.e. they form a vector with a length of K, where K represents the number of targeted classes (in our case, $K = 15$). A well-classified segment will therefore be characterized by a clear dominance (high proportion) of a class over the others, while a poorly classified segment will display a label with uniformly distributed proportion of classes. The tool is trained and confronted against the expert-based mapping classification of the Maupiti reef-lagoon system, which provided a single-class label for each part of the image. The models automatically process the same class at different depths, learning to differentiate between classes without considering the depth variation.

Our strategy is to efficiently combine both object-based and pixel-based approaches. The object-based classification is based on the segmented data (Section 5.2.3). Within each segment, statistical moments are computed for each of the four spectral bands, representing the first part of the explanatory variables of the dataset. The second part of the variables is provided by the pixel-based analysis. Based on the reflectance values from a small square centered on each pixel processed by a Random Forest (RF) classifier, the pixel-based classification assigns a specific class to each pixel (Section 5.2.3). The pixel-based features correspond to the number of pixels predicted to belong to each class within the considered segment. Finally, to obtain a visualizable map from the label data, we assign to each segment its majority class. Subsequently, we introduce a post-processing step in the workflow to

correct some misclassified segments. Note that two contrasted satellite sources are used to test the robustness and versatility of the mapping tool (Section 5.2.3).

Besides, the large size of the training dataset (encompassing millions of pixels) inherently results in the creation of deep decision trees within the RF models. Consequently, these models tend to be quite large, often exceeding several gigabytes in size. To address this, for both the pixel-based and object-based models, we developed two models: a full model and a lighter version characterized by fewer trees and shallower tree depth in comparison to the full model.

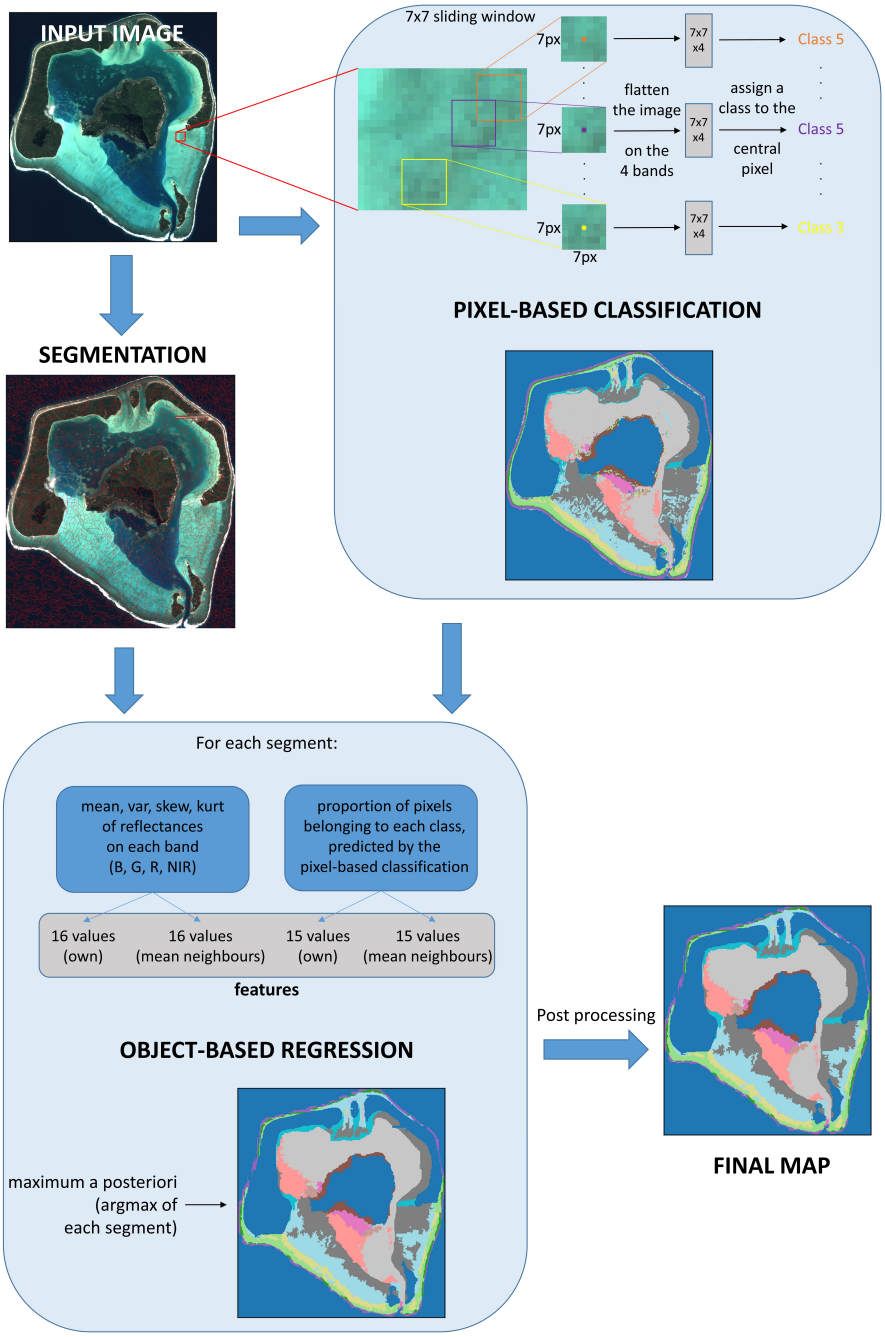


Figure 5.2: Full pipeline of the mapping tool.

Data sources

Our mapping tool is applied on two contrasted satellite sources: Sentinel-2 with high resolution (10m) and Pleiades [144], with very high resolution (2.8m). The Pleiades image offers exceptional resolution but is costly to obtain, limiting the size of the training sample for the model (here a single image is used). Additionally, its high resolution results in larger image sizes and longer processing times. By contrast, Sentinel-2 images are smaller, facilitating faster processing, which is a clear advantage during model development, and more easily accessible, allowing a more robust model training over a large database. For a given application, the final choice for the image source should therefore be dictated by user expectations and resources.

The Pleiades image is dated 14 June 2021, with minimal cloud coverage (see Figure 5.3). The size of the image is approximately 3900×3500 pixels. Four bands are available, all spanning a 2.8m resolution: blue, green, red, near-infrared (NIR).

Figure 5.3: Pleiades image of Maupiti island, 14 June 2021.

Our second data source comprises 8 images from the freely accessible Sentinel-2 satellite. These images were captured between 2020 and 2023 on selected dates with a low cloud cover. The Sentinel-2 resolution is 10 m for the blue, green, red, and NIR bands. Other bands are available, but have been discarded owing to their lower resolution (20m or 60m). The size of a Sentinel-2 image of Maupiti lagoon is approximately 700×700 pixels.

Pixel-based classification

The aim of the pixel-based classification is to predict the class to each pixel of the image. The classification accuracy is improved by using the information provided by neighboring pixels [152], [420]. An image is therefore created by selecting a square of $X \times X$ pixels around the pixel of interest, which creates an image of size $X \times X \times N$, having as a label the class of the central pixel (with N the number of bands, here $N = 4$). The pixels on the image edges are filled by mirroring the image. The image is then flattened to represent it as a vector of dimension X^2N, and fed to a RF classifier. Other classifiers were tested (Neural Networks,

Support Vector Machine, Decision Tree) but the RF classifier systematically yielded the best results. For each input, the RF classifier will return a single class which is the predicted class of the center pixel.

Pleiades image

For Pleiades image, X is fixed at 7 (square side about 20m). The RF classifier is trained with 5% of the pixels, which is enough to train the model owing to the large size of the Pleiades image. The classifier finally predicts the label of all the other pixels of the image. To avoid stocking all the $7 \times 7 \times 4$ images at once (which would lead to a $3900 \times 3500 \times 7 \times 7 \times 4$ matrix in the case of the Pleiades image), the predictions are made sequentially k rows by k rows, k being a number specified by the user (specified at 100 in our case).

Sentinel-2 images

For Sentinel-2 images image, X is fixed at 5 (square side about 50m) because of the lower image resolution. Using larger square with Sentinel-2 did not increase the performance but decreases the spatial accuracy, while using lower X value yielded inferior performances.

In addition, despite being normalized (level 2A), the Sentinel-2 images exhibit distinct reflectance values that are non-uniform and non-comparable. We performed a mean centering and variability reduction, that partially solved this issue. The model is then trained with 8 different images to capture more variability. It is also worth noting that a cloud-free image may not always be obtainable. To tackle this issue, users can create a mosaic image, i.e. an image which is a combination of several other cloudy images to obtain an almost cloud-free image. To take this case into account, one of the 8 training images is a mosaic image composed of three images (different from the 7 other training images). The 8 training images are finally displayed in Figure 5.4.

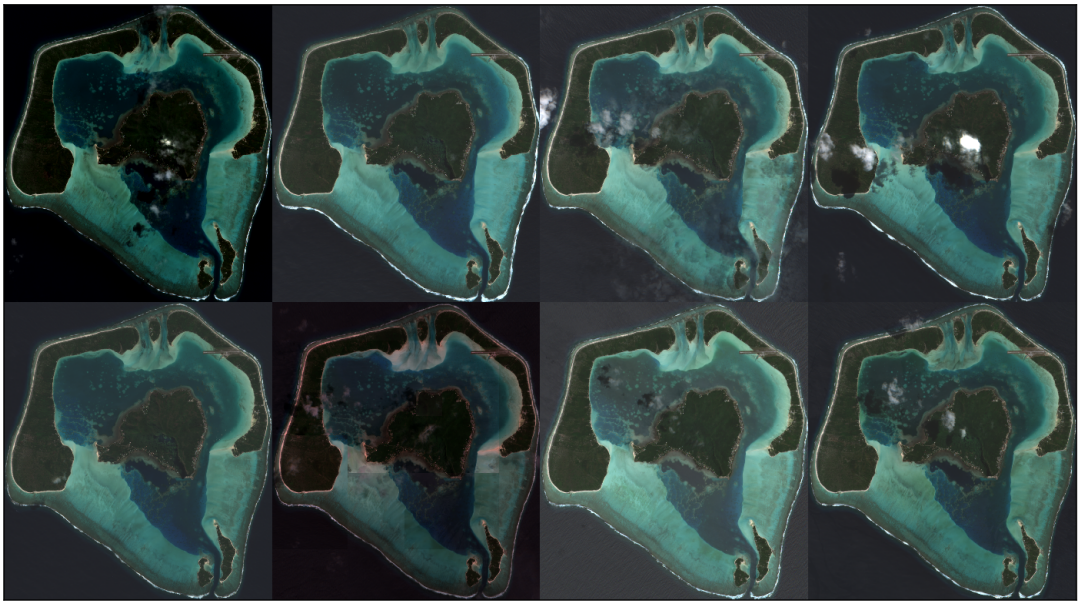


Figure 5.4: Raw Sentinel-2 images of Maupiti island used for training. The sixth image is a mosaic image.

Object-based segments regression

Segmentation

The image segmentation is performed using Felzenszwalb's segmentation method [117]. Other techniques were tested but lead to lower performance, see Section 5.4. As the segmentation only takes 3 bands over 4 as an input, the four possible combinations were tested out. The best performances were obtained with blue, green and red bands, i.e. the NIR band was abandoned for the segmentation process. An example of the segmentation is shown in Figure 5.5. A series of tests were performed to define the minimum segment size, i.e. the minimum number of pixels they contain, which can be set through a parameter of the algorithm. Based on expert-based observations, a large minimum size results on faster computation but lower accuracy, while small sizes lead to the creation of more segments and slowed the process. Optimized values are set at 20 and 200 for Sentinel-2 and Pleiades images, respectively, below which the accuracy was not further improved. Such minimal segment size correspond to areas between 1600 and 2000 m^2, which is in line with the expected representativeness for geomorphological elements.

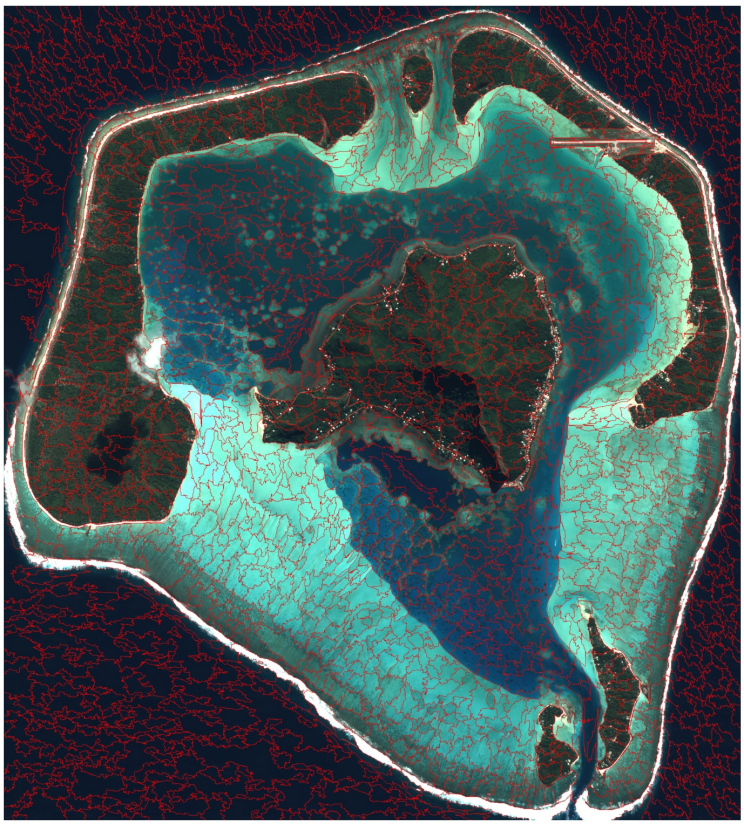


Figure 5.5: Example of Felzenszwalb's segmentation applied to the Pleiades image of Maupiti island.

Regression

The objects to be classified are the segments, i.e. the groups of pixels previously created with the segmentation method. Each of them is the statistical unit (i.e., row of the dataset), and the final length of the dataset will thus be the number of segments created.

For each segment, the explanatory variables are originating from two sources. The first source relies on the four statistical moments, i.e. the mean, variance, skewness and kurtosis of the reflectance values on each band, computed over each segment. These four statistical moments are chosen because they offer distinct and complementary information on the reflectance distribution (independence and non-correlation tests were performed using Principal Component Analysis). The four statistical moments computed over 4 bands provide 16 variables. To include the spatial context in the analysis (e.g. a segment surrounded by sand is most likely sand as well), the spatially-averaged value of these 16 variables over each neighboring segment is added to the dataset, resulting in a total of 32 variables. Note that neighboring segments are here simply defined as the segments sharing a border.

The second source of explanatory variables includes the results of the pixel-based classification. Within each segment, we enumerate the number of pixels that have been predicted (by the pixel-based classification) to belong to each class. A data vector of length K is then produced for each segment, where the j-th value is the proportion of the pixels belonging to class j. By definition, this is a compositional data [5]. Similarly to the statistical moments, the spatial context is accounted for by including the spatially-averaged values of the neighbouring segments, which adds $2 \times K = 30$ explanatory variables to the dataset, for a total of 62 variables.

Finally, during the training phase, we create the labels of this dataset (the target variable) to train the model. We want to obtain, for each segment, the true proportion of pixels belonging to each class. We hence perform the task of enumerating within each segment the number of pixels truly belonging to each class, from the ground-truth expert-based zonation. This gives a compositional vector of length K.

The shape of the labels, being compositional, is very special as most problems often focus on single-class classification or regression, or multi-output not being compositional. In our case, we decided to keep this shape because we were interested in the proportions of each class in each segment. One could argue that the proportion could also be obtained by the probability vector resulting in most models, however the accuracy of this output is higher when the model is initially trained with the compositional label.

The model selected is a RF regressor, as it was the one giving the best results and is often the best suited for coral mapping [298]. We proceeded to training and testing with a 5-fold cross-validation repeated over 10 times.

The object-based regressor provides a compositional vector label for each segment, which holds valuable information in its own right. However, to create a visual representation of the results at this stage, we take the argmax of this vector, (i.e. we use a *maximum a posteriori* decision rule). This operation allows us to assign each segment to the predominant class within it, enabling us to generate a visualizable map of the data.

Post-processing

At the end of the workflow, we introduce a final step aimed at refining the generated map through post-processing. The core concept behind this step is to rectify potential misclassifications. The approach is to compare each segment with all of its neighbors. If all the neighbors belong to the same class, we assign that class to the segment in question.

This post-processing strategy has demonstrated its effectiveness for relatively small segments where having an isolated segment of a particular class appears incongruous. However,

this approach may yield suboptimal results with larger segment sizes, as it becomes more plausible to encounter isolated segments that do not align with their neighbors.

Performance metrics

We define the *pixelwise score* of a mapping as the number of correctly predicted pixels divided by the total number of pixels. For the regression on the segments, the pixelwise score is computed after attributing to each pixel the majority class of the segment it belongs to. The *ideal pixelwise score* of a given segmentation is the score obtained when using the expert-based mapping. If the segmentation perfectly extract the same patterns than the expert, then the ideal pixelwise score is 1. In practice, this score is more often around 0.9. To remove this segmentation effect in the model evaluation, we can define the *scaled pixelwise score* as the score divided by the ideal pixelwise score. A high scaled score means that the mapping would be highly accurate if the segmentation was good. The pixelwise score is always computed after excluding land and ocean classes, because they are easy to predict and can represent a large amount of the dataset, thus introducing a bias in the score. Pixels used for training the regressor are excluded during the score computation.

Two metrics are used to compute the prediction accuracy for compositional data labels. The first one is the coefficient of determination R^2, which can be computed on each class and then averaged. By denoting y_{ij} the real value of a sample i for a class j, and $\hat{y}_{ij}$ its predicted value, the coefficient of determination for the class j is

$$R_j^2 = 1 - \frac{\sum_{i=1}^{n}(y_{ij} - \hat{y}_{ij})^2}{\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)^2},$$

where $\bar{y}_j$ represents the mean of the values in the jth class.

Another suitable metric to compare the distance between two compositional labels is the cosine similarity [294], [389], [454], representing the cosine of the angle formed between the two vectors,

$$\text{Cos similarity} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i \hat{y}_i^\top}{\|y_i\|_2 \cdot \|\hat{y}_i\|_2},$$

where y_i (resp. $\hat{y}_i$) is the compositional vector of all the classes for the sample i (resp. its predicted value).

Evaluation strategy

Both training and testing sets are defined for each image source. For Sentinel-2, the training set is the 8 training images displayed in Figure 5.4 and the testing set are two unseen images (see Section 5.3.2). For the single Pleiades image, the training set consists of randomly selecting 80% of the segments, and the testing set is the remaining 20%. The model is evaluated on three different levels.

- Pixel-based only: the pixel-based model training against the whole training set being computationally prohibitive, the training is carried out using a subset of the training set. For both Pleiades and Sentinel-2, the subset consists in selecting a maximum of 50,000 of each class, acknowledging that some smaller classes have a maximum of only 10,000 points. Note that these points are only selected among the training data, so the testing data only consists of unseen pixels and segments. Then, a map of the whole testing set is predicted and the metrics are measured on it.

- Object-based: the object-based model is trained on the full training set, and then used to predict the labels for each segment over the test image. For each segment, the *maximum a posteriori* is computed and we assign a single class to the segment. Similarly as for the pixel-based, for Pleiades the metrics are only measured on the testing set.

- Post-processing: this steps corresponds to the final, post-processed map.

We also performed hyperparameter tuning to find the optimal parameters of all the models.

5.3 Results

5.3.1 Expert-based Maupiti mapping

The expert-based zonation produces the map displayed in Figure 5.6. On the square center on the island, approximately 50% of the pixels are either ocean or land and are thus unclassified (white color). Among the remaining 50%, the classes are highly unbalanced, with more sand than coral. The spatial distribution of the area covered by each class is given in Table 5.1. In more details, one notes the cross-shore gradient in seabed type at both main barrier reefs (South-West and East) with, from the forereef to the inner lagoon, the nearly systematic succession of C-SAG, C-FCO, C-MCD, S-MSC, S-SSC and S-DS. The S-SS class develops in areas of sediment convergence, at both sides of the central island and at the nothern deposit. C-SHR class is, by definition, observed around the central island, while C-RER displays two distinct areas west and southwest of the central island. The shallow part of the northern breaches displays a mixing of S-MSC, C-MCD and C-AER.

Table 5.1: Area covered by each class.

	Coral superclass									Sand superclass				
Class	AER	CDR	FCO	MCD	RER	SAG	SHR	TSR	DW	DS	LSC	MSC	SS	SSC
Area (%)	0.5	0.1	2.7	6.3	6.4	4.9	3.1	1.5	1.7	19.1	31.1	4.6	5.8	12.1

5.3.2 Automated mapping tool

Pleiades results

The hyperparameters used to train the models are presented in Appendix 5.A. For this single image data, the results are computed based on a 5-fold cross-validation strategy, with 10 repetitions. The training subset, created by selecting a maximum of 50,000 points of each class, spans approximately 5% of the total dataset.

For the pixel-based model, after hyperparameter tuning, the full model has 200 trees with depths of up to 40 and the light model has 50 trees with maximum depth of 15. The full model occupies 3.58 gigabytes (GB) of storage space, while the lighter model is significantly smaller at 92 megabytes (MB). In contrast, the object-based model is already relatively compact, utilizing only 50MB of storage space with 200 trees and a maximum depth of 20. However, due to its dependency on the outputs of the pixel-based model, we trained two

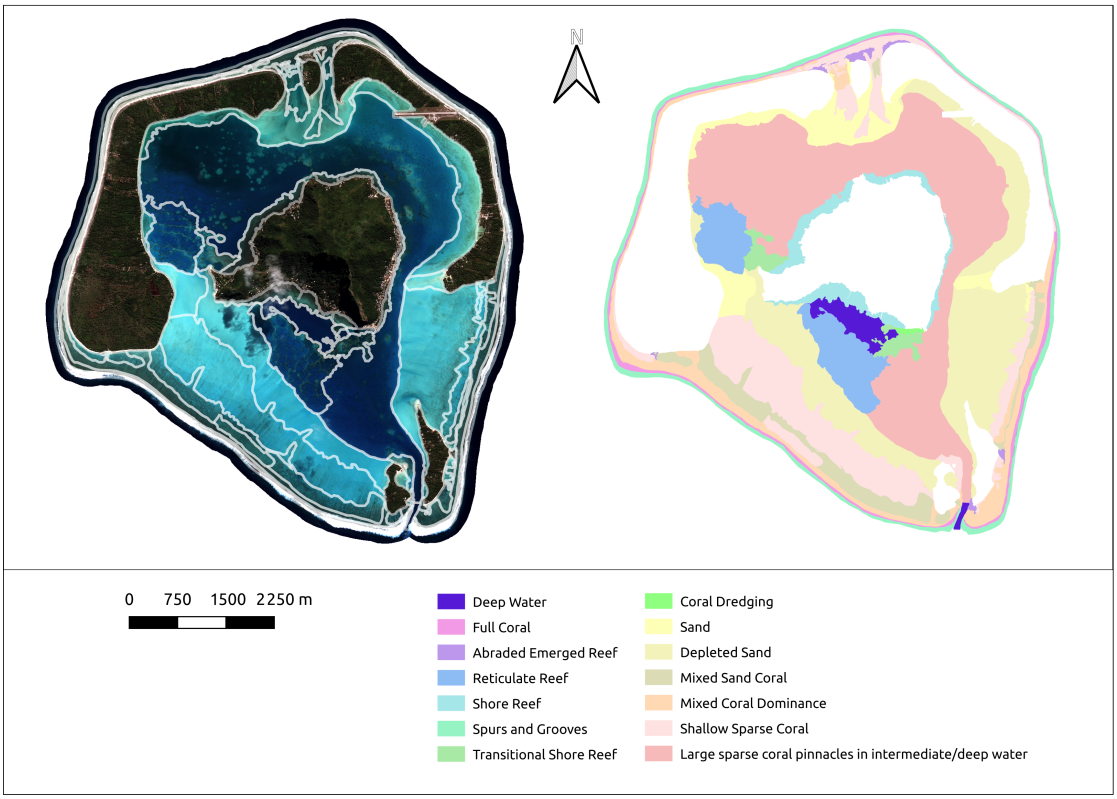


Figure 5.6: Expert-based zonation of seabed classes. Right: satellite view and zone delimitation. Left: complete expert-based zonation.

versions of the object-based model. Each of these versions utilizes either the output of the full pixel-based model or the light pixel-based model.

Despite being less accurate, the lighter model serves two crucial purposes. First, it facilitates easy distribution to users, ensuring that they can efficiently access and use the model without the burden of handling a massive file (the light model is below the GitHub file size limit of 100MB). Second, it significantly speeds up loading times, enhancing the overall user experience by reducing the time required to deploy the model for analysis. The performances for the full (resp. light) model are presented in Table 5.2 (resp. Table 5.3).

Table 5.2: Mean performances of the full Pleiades model with a 5-fold cross-validation repeated 10 times. The standard deviation appears within parenthesis.

Model	Pixelwise	Scaled pixelwise	R^2	Cos sim.
Pixel-based only	0.8208 (0.0073)	-	0.6765 (0.0203)	0.9551 (0.0008)
Object-based	0.8636 (0.0083)	0.9369 (0.008)	0.9217 (0.0008)	0.9916 (0.0004)
Post-processing	0.8659 (0.0086)	0.9394 (0.0079)	-	-

In these tables, the metrics are not only applicable to all the models. For instance, the pixel-based model does not depend on a segmentation, so it does not have any scaled pixelwise score. The R^2 and cosine similarity metrics are computed on compositional vector labels, but the post-processing step only has a single class label associated to each segment. Because of that, it cannot have these metrics.

Table 5.3: Mean performances of the light Pleiades model with a 5-fold cross-validation repeated 10 times. The standard deviation appears within parenthesis.

Model	Pixelwise	Scaled pixelwise	R^2	Cos sim.
Pixel-based only	0.7616 (0.0091)	-	0.3831 (0.03)	0.8887 (0.0071)
Object-based	0.849 (0.0079)	0.9211 (0.0077)	0.8887 (0.0071)	0.9862 (0.0004)
Post-processing	0.8534 (0.0083)	0.9258 (0.0076)	-	-

For both models, the performances measured on all the metrics are improved across the different steps of the workflow. For the full model, the pixelwise score goes from 82% for the pixel-based classifier to 86% for the object-based regressor. The improvement is higher with the light model, where the pixelwise score goes from 76% to 85%. The R^2 also shows an important improvement, from 0.68 to 0.92 for the full model, and from 0.38 to 0.89 for the light model.

When comparing the performances between the full and light model, there are no significant differences in the final performance between the full and light models. The only noticeable distinction emerges when examining the pixel-based classifier, where a 3% difference between the two models is observed. It is worth noting that the object-based regressor does not show any degradation in performance between the two model variants.

Sentinel-2 results

The hyperparameters used to train the models are presented in Appendix 5.A. As we work with various Sentinel-2 images, we are able to evaluate the models on different datasets. We conducted performance tests on two distinct images for a comprehensive assessment:

1. A good quality image dated August 23, 2023, free from any cloud cover.
2. A mosaic image created by combining several cloudy images dating between 2021 and 2023. We assembled this mosaic in a way that minimizes cloud cover. Assessing performance on this second image is essential as it may represent the only available dataset for some users.

Similarly as Pleiades data, a full and a light model are built. The full pixel-based model is 3.15GB, and the light model is 97MB. For the object-based model, the full version is 298MB and the light one 87MB. We present the performance results of the full (resp. light) model in Table 6.1 (resp. Table 5.5).

Table 5.4: Performances of the full Sentinel-2 model. The ideal pixelwise score is indicated within parenthesis below the image name.

Validation image	Model	Pixelwise	Scaled pixelwise	R^2	Cos sim.
23-08-2023 image (Ideal = 0.9051)	Pixel-based only	0.7955	-	0.7000	0.9315
	Object-based	0.8123	0.8974	0.7933	0.9541
	Post-processing	0.8130	0.8983	-	-
Mosaic image (Ideal = 0.9043)	Pixel-based only	0.7529	-	0.6296	0.9132
	Object-based	0.7934	0.8774	0.7256	0.9415
	Post-processing	0.7942	0.8782	-	-

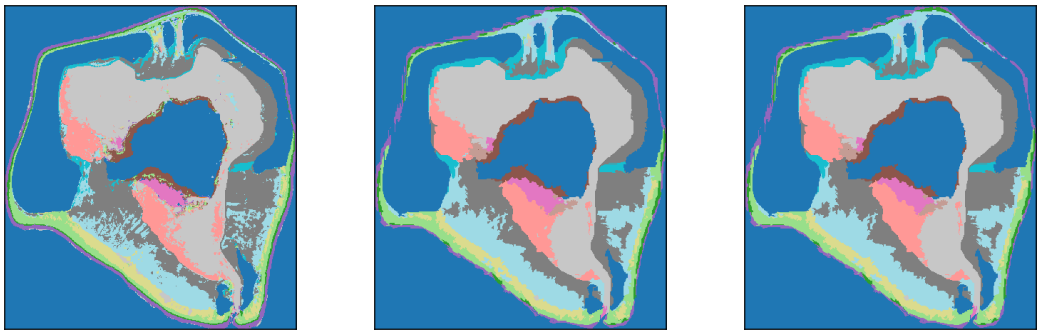
Similarly to the Pleiades image, the performances of the full and light model are similar, excepted when examining the pixel-based classifier only which shows a slight difference of 3% for the cloudfree image and 2% for the mosaic image.

Table 5.5: Performances of the light Sentinel-2 model. The ideal pixelwise score is indicated within parenthesis below the image name.

Validation image	Model	Pixelwise	Scaled pixelwise	R^2	Cos sim.
23-08-2023 image (Ideal = 0.9051)	Pixel-based only	0.7625	-	0.6204	0.9143
	Object-based	0.8086	0.8934	0.7712	0.9509
	Post-processing	0.8095	0.8943	-	-
Mosaic image (Ideal = 0.9043)	Pixel-based only	0.7308	-	0.5641	0.8991
	Object-based	0.7962	0.8804	0.7115	0.9416
	Post-processing	0.7970	0.8813	-	-

Comparing the pixelwise score on the good quality, cloud-free image with that on the mosaic image logically reveals better results for the former. However, the difference is relatively small, with only a 2% decrease for the full model and a 1% decrease for the light model. This demonstrates that the model has effectively learned to handle variability during the training phase, enabling it to generate accurate maps from images with different levels of inherent variation. The primary variation between the two types of images is in the R^2 score, which exhibits a 7% difference.

At each of the three steps (pixel-based, object-based, post-processing), it is possible to create a map. Figure 5.7 displays the three maps returned by the model with the 23-08-2023 satellite image.



(a) Pixel-based map (b) Object-based map (c) Post-processed final map

Figure 5.7: Maps created by the model with the 2023-08-23 Sentinel-2 image as an input.

5.4 Discussion and conclusion

Monitoring coral reefs in order to assess their reactions to environmental changes is both a crucial need for coral habitat protection and restoration and a significant scientific challenge [16], [216]. To be practically operable, monitoring study rely on seabed mapping, i.e. the assignation of specific classes to distinct areas of the considered reef system. The first necessary step is therefore to define an ensemble of the specific classes that must be distinguished. The strategy used here is to base the classification on fine topographical features of the seabed, which can be considered as a variation of usual benthic classifications [151]. The aim of such topography-based classification is to map and to monitor the geometrical structure of seabeds, which is of primary interest for habitat surveys, parameterization of wave-circulation numerical models and sedimentary and morphodynamical studies. A specific set of classes has been therefore proposed to establish an expert-led mapping of the study site from

field and high-resolution satellite date. Due to the permanent adaptation of biological and geological processes to meteo-marine forcing exposure [101], the obtained mapping in Figure 5.6 has naturally significant matching with existing geomorphic classification [13]. Therefore, partial similarities can be found between the present classes and more traditional geomorphic compartments [216]. However, it should be emphasized that the variability of seabed architecture observed on the Maupiti study site can not systematically be described using geomorphic classes. Moreover, as a complement to traditional geomorphic and habitat-related approaches, the longer term aim of the present project is to build a quantitative description of seabed architecture, i.e. based on measurable metrics of the high-resolution (sub-metric) seabed elevation. This work will be fed by ongoing and future high-resolution surveys.

The second crucial step of coral reef monitoring is the development of automated tools on satellite imagery, targeting the scale convergence between downscaling of remote sensing and upscaling of in-situ observations [216], [298]. A key point of the present study is the comparative application of the mapping tool on two different image sources aiming to assess the versatility of the proposed approach. The two present models designed for Pleiades and Sentinel-2 serve distinct purposes and are evaluated on different types of data. Their performances should be interpreted in light of these differences. The Pleiades model's performance serves as a base benchmark for the proposed methodology. This model is trained on exceptionally high-resolution imagery (2.8 meters) and undergoes a 5-fold cross-validation on the same image. However, it is worthwhile to note that this approach may not fully capture the inherent variability present in all satellite images. Even with corrections applied, images of the same location captured at different times can exhibit variations in reflectance values [251]. Therefore, when assessing the high performance of the Pleiades model, which achieves over 86% pixelwise accuracy and a R^2 score of 0.92, it remains essential to consider the testing set which does not vary from the training set, thus making it easier to perform well. On the other hand, the performance of the Sentinel-2 model provides a more realistic view of its usability in practical scenarios, given that Sentinel-2 has been widely used in the recent years for mapping purposes [298], [322]. This model is evaluated on two unseen images, representing the challenges users may encounter: either a good quality, cloud-free image from August 23, 2023, or a mosaic image constructed to mitigate cloud cover (along with the associated variability). In these real-world scenarios, the accuracy slightly decreases to a range between 79% and 81%, with the R^2 score dropping to less than 0.8.

The study evaluated the proposed trained workflow on two Sentinel-2 images, reflecting two real-world scenarios users may encounter: either using a good quality, cloud-free image, or a mosaic image reconstructed to mitigate cloud cover. The major difference in performance between the two images comes from the R^2 score. However, this score represents the model's ability to predict compositional labels, which may be less critical for users solely interested in obtaining a geographical map of a location. While it remains advisable to use high-quality, cloud-free images when available, the model maintains consistent performance even with mosaic images, ensuring its reliability across diverse real-world scenarios [129], [248].

For each satellite model, we have introduced two variants: a full model and a lighter one. The full model acts as a reference point, providing insights into the potential performance of our methodology. However, distributing this full model presents challenges due to its substantial size, weighing several gigabytes. Consequently, we have developed a lighter model

with equal performance, ensuring practical usability. The only difference between the two models was found on the pixel-based, which showed a slightly better performances for the full model. This may be attributed to the fact that the most important reduction in size between the full and light models is found within the pixel-based classifier. Indeed, its light version is 30 times smaller than its full counterpart, inevitably impacting its performance to some extent [332], [373]. On the other hand, the object-based regressor has the same size for both models, and does not show any difference in performances.

Moreover, the workflow proposed in this paper has been applied to coral reef mapping, but its structure can be adapted beyond this scope. Our work was thought so that the proposed tool is adaptable and can be retrained to handle different types of data. For instance, the workflow has been tested with geomorphic and benthic data of the Great Barrier Reef from the Allen Coral Atlas project [32], by considering their maps as expert-based maps. The results were consistent, although not publishable due to the nature of the Allen Coral Atlas maps, which are already generated by machine learning models, thereby lacking the rigor associated with a ground-truth map. A significant limitation to the retraining process is that it requires access to a full ground-truth map, which can be challenging to obtain. Nevertheless, even when only a limited number of ground-truth points are available, users can still use the first part of the workflow. Indeed, the pixel-based classification, which relies on sparse points (although a substantial quantity of them), can be executed independently. One possible approach for a user who lacks a full ground-truth map is to conduct both the segmentation and the generation of the pixel-based map. Afterwards, for each segment and each class, users can count the number of predicted pixels belonging to each class within that segment. Subsequently, the majority class can be assigned to all the pixels in that segment. We refer to this process as "smoothing" [311] and give an example in Appendix 5.B.

Furthermore, the use of different segmentation techniques has not been thoroughly addressed in this work. We see that the optimal pixelwise score, which assesses how well a segmentation aligns with the expert-based mapping, frequently hovers around 90%. Essentially, this implies that the highest achievable accuracy for any model was 90%. In practice, our model for Sentinel-2 data achieved approximately 80% accuracy, equivalent to about 89% when scaled against the best possible performance. During the development of this tool, we explored several segmentation methods, including Felsenszwalb's method [117] as well as Quickshift [427], SLIC [1] and compact wathershed [296]. Among these options, Felsenszwalb's method delivered the best results, that is why we kept it in our tool. However, it's worth noting that all of these methods are over a decade old, and we did not delve deeper into exploring newer alternatives. Hence, there is a promising avenue for improvement by incorporating state-of-the-art segmentation algorithms into our workflow. To illustrate the potential impact, imagine we find a segmentation technique achieving a score of 95%. If we assume that our scaled score remains consistent, this enhancement could elevate the final tool's accuracy to almost 85%. Therefore, exploring and implementing an advanced segmentation algorithm holds the potential for a substantial performance boost.

In conclusion, we have introduced a workflow capable of delivering highly accurate topography-based seabed mappings derived from satellite imagery. We tested its performance on two satellite sources with varying resolutions to demonstrate its effectiveness across diverse images. Moreover, our workflow is adaptable, allowing for retraining on various datasets and providing maps for different sources, and easily testable on other types of seabed

classifications. The provided tool is pre-trained and readily applicable for use with Sentinel-2 images.

Appendix 5.A Hyperparameters of the models

The models for the Sentinel-2 data are trained with the *scikit-learn* Python library. The pixel-based model is a RandomForestClassifier and the object-based is a RandomForestRegressor. The hyperparameters used for both of them are indicated in Table 5.6 for the Pleiades model, and Table 5.7 for the Sentinel-2 model. The arguments *max_depth* and *n_estimators*, respectively representing the maximum depth of a tree and the number of trees, depend on whether the full or light version is used.

Table 5.6: Hyperparameters of the RF models for the Pleiades data. The random_state argument is 0 for both of them.

Hyperparameter name	Pixel-based RF	Object-based RF
criterion	'gini'	'squared_error'
max_features	'sqrt'	'sqrt'
min_impurity_decrease	0	0
min_samples_leaf	1	1
min_samples_split	4	5
max_depth (full)	40	20
n_estimators (full)	200	200
max_depth (light)	15	15
n_estimators (light)	50	80

Table 5.7: Hyperparameters of the RF models for the Sentinel-2 data. The random_state argument is 0 for both of them.

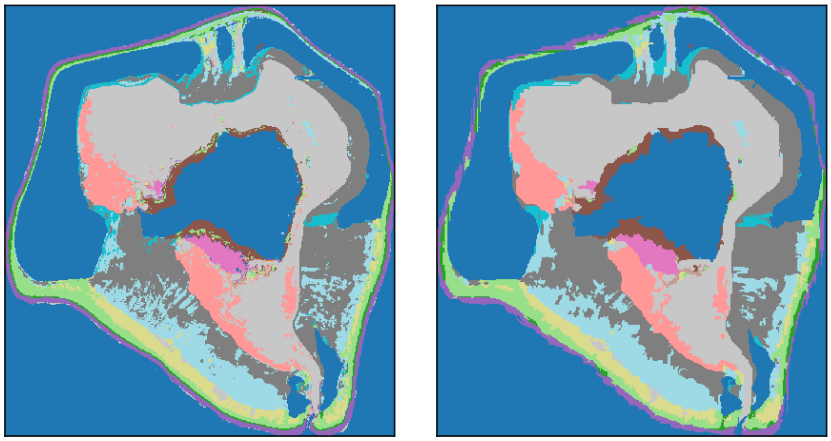
Hyperparameter name	Pixel-based RF	Object-based RF
criterion	'gini'	'squared_error'
max_features	'sqrt'	'sqrt'
min_impurity_decrease	0	0
min_samples_leaf	1	1
min_samples_split	4	5
max_depth (full)	50	20
n_estimators (full)	400	300
max_depth (light)	15	20
n_estimators (light)	50	80

Appendix 5.B Performances of the smoothed pixel-based model

We take the example of the prediction on the 23-08-2023 satellite image. We first predict the pixel-based map (Figure 5.B.1a). Then, we proceed to the segmentation and each segment is assigned to the majority predicted class of its pixels. This gives a map that looks smoother than the pixel-based one (Figure 5.B.1b) because it erases the individual pixels by grouping them with their neighbours. By stopping at this step, it is possible to train the pixel-based model only with sparse ground-truth points.

In Tables 5.8 and 5.9, we compare the performance of the smoothed pixel-based map to that of other maps when utilizing the full model on both Pleiades and Sentinel-2 images.

For Pleiades imagery, the performances of the smoothed pixel-based map are higher than the performances of the pixel-based map, but lower than the object-based map. In



(a) Pixel-based map (b) Smoothed pixel-based map

Figure 5.B.1: Pixel-based map and smoothed pixel-based map with the 2023-08-23 Sentinel-2 image.

Table 5.8: Mean performances of the full Pleiades model with a 5-fold cross-validation repeated 10 times. The standard deviation appears within parenthesis.

Model	Pixelwise	Scaled pixelwise
Pixel-based only	0.8208 (0.0073)	-
Smoothed pixel-based	0.8409 (0.0076)	0.9123 (0.0082)
Object-based	0.8636 (0.0083)	0.9369 (0.008)
Post-processing	0.8659 (0.0086)	0.9394 (0.0079)

Table 5.9: Performances of the full Sentinel-2 model. The ideal pixelwise score is indicated within parenthesis below the image name.

Validation image	Model	Pixelwise	Scaled pixelwise
23-08-2023 image (Ideal = 0.9051)	Pixel-based only	0.7955	-
	Smoothed pixel-based	0.7760	0.8574
	Object-based	0.8123	0.8974
	Post-processing	0.8130	0.8983
Mosaic image (Ideal = 0.9043)	Pixel-based only	0.7529	-
	Smoothed pixel-based	0.7486	0.8278
	Object-based	0.7934	0.8774
	Post-processing	0.7942	0.8782

contrast, when applied to Sentinel-2 imagery, the performance of the smoothed pixel-based map is lower than that of the original pixel-based map. This discrepancy arises because some correctly classified pixels are reassigned to an incorrect class due to the misclassification of other pixels within the same segment.

Our observation suggests that the smoothing step is particularly effective with very high-resolution images like Pleiades, where it improves overall performance. However, its efficacy may diminish when applied to lower-resolution images, such as Sentinel-2, where the trade-off between improving misclassified pixels and potentially affecting correctly classified ones becomes more pronounced.

Conclusion

This chapter introduced the workflow developed throughout the course of this thesis. The final tool was trained and tested using data from two distinct satellite sources: Pleiades and Sentinel-2, each characterized by varying spatial resolutions. The workflow consistently delivered strong performance for both sources, achieving an accuracy exceeding 85% across 15 distinct classes.

The workflow utilizes supervised Random Forest models and combines pixel-based and object-based methodologies. Additionally, to account for spatial dependencies, it incorporates the spatial information by computing the averages of explanatory variables from neighbouring points.

As it stands, the tool is fully trained and readily operational for use with Sentinel-2 satellite images. Moreover, the models are adaptable and can be re-trained as necessary to address the specific requirements of users, thereby extending the tool's utility to a wide array of applications, for instance the classification of forests or ice sheets.

CHAPTER 6

Conclusion

6.1 Context

In the midst of the global climate change crisis and the alarming decline in biodiversity, it has become imperative to study species and gain insights into their locations and temporal dynamics. A pivotal tool for achieving this is the creation of presence maps that capture various attributes of these species.

This thesis places its focus on coral reefs, with a specific case study centered around Maupiti island in French Polynesia. Gathering data on coral reefs can be an arduous task for several reasons. Firstly, the sheer expanse of these ecosystems, such as the Great Barrier Reef spanning 348,000 square kilometers, often exceeds the coverage of human expeditions. Secondly, many coral reefs are situated in remote and less accessible locations, exemplified by the case of French Polynesia. To address these challenges, we turned our attention to the potential of satellite imagery as a valuable resource for mapping coral reefs.

Given that manual mapping of such expansive areas can be exceedingly labor-intensive, this thesis delves into the realm of machine learning algorithms as a solution. What sets our expert-based data apart is the availability of a full ground-truth map, in contrast to sparse point-based references. This attribute opened doors to employ methods that would otherwise be impractical.

The development of this mapping tool brought to light two challenges: the intricate nature of compositional data, a distinctive type of data that conveys relative information; and the spatiality in the dataset. Compositional data appeared during the segmentation of satellite images, where pixels are grouped based on shared features. Spatiality naturally emerged from the fundamental handling of satellite imagery. The primary objective of this thesis was to discern the most effective techniques for handling the compositionality and spatial characteristics within this data. Furthermore, we aimed to identify the models best suited to process this compositional data, with the ultimate goal of generating satellite-derived coral reef maps.

6.2 Contributions

This thesis introduced an automated satellite-based mapping tool capable of generating coral reef maps with an 85% accuracy across 15 distinct classes. Beyond this tool, which is readily applicable to Sentinel-2 and Pleiades images, the proposed workflow can be easily replicated by anyone with access to the necessary training data (a full ground-truth map). The research

unfolded through four distinct studies.

Literature review

The first study was a literature review of the scientific papers published between 2018 and 2020 that focused on multispectral satellite-based coral classification. This review served as the foundational step in our research, providing valuable insights into multispectral satellite imagery, preprocessing techniques, pixel-based and object-based methods, as well as the most efficient machine learning models.

This chapter showed the superiority of two models: the Random Forest (RF) and the Neural Network (NN). In the final mapping tool, we made the choice to only use the RF for both the pixel-based and the object-based models. This choice was driven by the better performances of the RF after trying both of them.

Besides, this first study highlighted that preprocessing algorithms globally improved model accuracy. However, in our tool, it appears that we did not employ these preprocessing algorithms. Below, we provide a breakdown of each preprocessing technique mentioned in the same order as Chapter 2, explaining whether we utilized it and the reason behind our choice.

1. **Clouds and cloud shadows**: The Pleiades image we employed was nearly cloud-free, except for a small cloud on land. For Sentinel-2 images, we carefully selected ones with relatively low cloud cover (<5%). We also generated a composite image from multi-temporal images to test the model's performance on this kind of images.

2. **Water-column correction**: We implemented Lyzenga's water-column correction algorithm [258] during the thesis and made it publicly available on GitHub (https://github.com/teongu/lyzenga1978). Although we did not subject it to rigorous cross-validation testing, initial results indicated only marginal performance enhancement. Additionally, the algorithm required prior knowledge of bathymetry, limiting its accessibility. Consequently, we decided not to include it in the final mapping tool.

3. **Light scattering**: This phenomenon was automatically corrected in the acquired satellite images.

4. **Masking**: Since we primarily used low-cloud-cover images, we estimated that cloud masking was unnecessary. For ocean and land classes, we retained them as separate categories and reduced their prevalence through undersampling to maintain a balanced dataset without overemphasizing these classes.

5. **Sunglint removal**: We developed Hedley's deglint method [176] with one of my interns. While a comprehensive evaluation was not conducted, initial results did not demonstrate significant performance improvements. To avoid increased computational time and user burden (requiring users to define a sunglint zone), we chose not to include it.

6. **Geometric and radiometric corrections**: Both the Pleiades image and Sentinel-2 level 2A images we utilized had already undergone these corrections.

7. **Contextual editing**: We implemented a simple post-processing step in which we examined each segment's neighbors. If all neighbors belonged to the same class, we

assigned that class to the segment. This strategy proved effective for relatively small segments but may yield suboptimal results with larger segment sizes. However, we did not conduct an in-depth post-processing editing, which would necessitate a deep understanding of reef patterns (which I currently lack). Therefore, the potential of contextual editing remains unexplored.

SMOTE for Compositional Data

The second study introduced SMOTE for Compositional Data (SMOTE-CD), an adapted oversampling method for compositional data based on the original SMOTE algorithm [75]. Compositional data are data carrying relative information, and they became a focal point during our tool's development. Indeed, each segment we worked with had labels representing proportions of various constituent classes.

As the need to balance dataset emerged to enhance the model's performance, we found no existing method in the literature to address this specific requirement. Consequently, we developed our own method to address this gap. While this technique proved effective for a limited number of classes, its efficiency diminished with a higher number of classes, rendering it unnecessary for our study, which involved 15 classes. Besides, even though the performance of the final mapping tool showed slight improvements with SMOTE-CD, we ultimately decided against its use. Indeed, the need for SMOTE-CD was deemed unnecessary because we already had a substantial number of training segments, exceeding 40,000. This quantity was more than enough for efficient model training. Besides, SMOTE-CD, as an oversampling technique, augmented the number of segments, hence slowing down the training process.

Consequently, we opted not to utilize SMOTE-CD in our workflow, considering the trade-off between its potential benefits and the significant increase in computational time it introduced.

SAR Dirichlet

Compositional data also played a pivotal role in the third study, where we incorporated spatial information into such data. Drawing upon the Dirichlet distribution, a naturally well-suited distribution for compositional data, we developed a Spatial AutoRegressive (SAR) model by introducing a spatial lag term to a Dirichlet regressor. Although this enhancement did indeed improve performance compared to a non-spatial Dirichlet model, we made decision not to include it in our final workflow for two reasons.

First, the SAR Dirichlet model is very slow, both for training and for prediction. Notably, the time cost of the model is in $O(n^3)$, with n the number of samples. Given that the model was starting to be slow to process with 10000 samples, it was not manageable to use it in the real framework where the number of segments can be hundreds of thousands. However, this first reason is mainly caused by the fact that we did not have enough time to tackle the matrix inversion problem that arises in the SAR model. Indeed, by avoiding to inverse the matrix or using more optimal methods, the computational time could be highly reduced [160], [312].

The second, and arguably more crucial, reason for not incorporating the SAR Dirichlet model into the final tool was its performance. While the model performed reasonably well

with a small number of classes (fewer than 5), its effectiveness diminished when dealing with a higher number of classes. In our case, which involves 15 classes, the SAR Dirichlet model delivered unsatisfactory results, and its accuracy fell significantly short of expectations. The SAR Dirichlet model was outperformed by other machine learning models, notably the RF. Consequently, the latter was chosen for our final workflow.

Mapping tool

In the last study, the final mapping tool is presented. The originality of the methodology lies in the ground-truth data, because rather than relying on a conventional sparse sampling, we possess the complete mapping of Maupiti lagoon. This distinctive advantage enabled us to develop a workflow integrating both pixel-based and object-based models, hence reaching an accuracy exceeding 85%.

Two models are proposed, trained with different satellite images having different resolution (2.8m for Pleiades, 10m for Sentinel-2). From this study only, we can not compare the performances of the two models because they are not tested on a fair ground: the Pleiades model is tested on the same image it is trained with, while the Sentinel-2 model is tested on different, unseen images, exposing it to the variability of the reflectances perceived by the sensor.

To avoid this bias, we also trained and tested a model on a single Sentinel-2 image (23rd August 2023), using a 5-fold cross-validation, to be able to compare it equally to the Pleiades model. The results, presented in Table 6.1, suggest that the performances of this Sentinel-2 model are equivalent to the Pleiades model presented in the last chapter. This is interesting as it means that an increase in resolution does not necessarily imply an increase in performances.

Table 6.1: Mean performances of the full Sentinel-2 model on a single image, with a 5-fold cross-validation repeated 10 times. The standard deviation appears within parenthesis.

Model	Pixelwise	Scaled pixelwise	R^2	Cos sim.
Pixel-based only	0.8282 (0.0106)	-	0.8877 (0.0111)	0.9769 (0.0008)
Smoothed pixel-based	0.8313 (0.014)	0.9112 (0.0132)	-	-
2nd round	0.8553 (0.012)	0.9376 (0.106)	0.9197 (0.0101)	0.9922 (0.0005)
2nd round + post-processing	0.8623 (0.0118)	0.9452 (0.0105)	-	-

We believe that there are two reasons for which a user might want to use a Sentinel-2 image instead of a high-resolution image like Pleiades. First, the price: Sentinel-2 images are free of access while high-resolution images are often quite expensive to acquire, especially if we want to cover a large area. Second, the computational speed: because of their difference in resolutions, it is between 10 and 15 times longer to map a Pleiades image compared to a Sentinel-2 image. This difference may be negligible when mapping a small area, that would take one hour instead of a few minutes, but it becomes problematic on larger areas. For instance, for an area of hundreds of kilometers square, the computation may take days instead of hours, which is not affordable by everyone.

Code distribution

Finally, a substantial portion of the work are available on GitHub https://github.com/teongu, along with tutorial to ease their application. These resources include the Lyzenga's water

column correction, the SMOTE-CD, the SAR Dirichlet and the final mapping tool. To enhance usability, we also developed a Python package for SMOTE-CD (https://smote-cd.readthedocs.io/en/latest/).

6.3 Perspectives

Spatial information

In Chapter 5, spatial information was integrated into the model by augmenting the feature set, effectively doubling its size, and including the mean values of neighboring features. This modification significantly enhanced the model's performance, representing a straightforward approach to address spatial considerations. Nevertheless, to achieve a more comprehensive and potentially higher-accuracy solution, the prospect of directly modifying the model itself (in this case, the RF) to incorporate spatial characteristics remains unexplored and worthy of investigation [137], [375].

Multispectral bands

Chapter 2 proposes a list of the most common satellites, many of which offer more than four spectral bands. For instance, the Sentinel-2 satellite has eight bands at its disposal. However, in our workflow, we limited our usage to just 4 specific bands (green, blue, red, near-infrared). This selection was primarily driven by the difference in spatial resolution among these bands, as the chosen four offer a 10m resolution, whereas the remaining bands feature 20m or 60m resolutions, rendering their practical application quite challenging.

Nevertheless, it's worth noting that our object-based regressor leverages statistical moments such as mean, variance, skewness, and kurtosis. This suggests that the additional information contained in those bands could potentially yield valuable insights, especially considering our focus on coral reefs, where certain blue bands can penetrate deeper into the water column. To harness these bands effectively, further efforts would be necessary, and we did not explore this possibility due to time constraint. Besides, adding more information would entail a computational cost in terms of speed, but it holds the promise of potentially enhancing overall performance.

Post processing

Earlier in this section, we discussed the post-processing step we implemented, which primarily involved assigning isolated segments the same class as their neighbors. However, there is further potential for enhancing the mapping tool through contextual editing. This entails refining the map using known patterns of coral reefs, which could substantially improve the accuracy and reliability of the tool [290].

We believe that an in-depth post-processing phase, guided by well-defined rules governing coral reef distribution, holds promise for improving the tool's performance. Alternatively, we could explore the development of a dedicated machine learning model designed specifically for this task. The array of possibilities in this domain remains extensive and uncharted by our work.

6.4 Overall conclusion

This thesis proposed a coral reef mapping tool combining both pixel-based and object-based methods. The synergy between these two techniques, rather than relying solely on one or the other, has allowed us to achieve an accuracy exceeding 85% across 15 distinct classes.

Notably, the tool's performance remains consistent irrespective of the satellite image resolution, whether it is Sentinel-2 (with a 10-meter resolution) or Pleiades (with a finer 2.8-meter resolution). This suggests that users who do not require ultra-high resolution can obtain maps with the same level of accuracy, all while benefiting from reduced computational costs due to the smaller image size.

An interesting observation is that our models do not necessarily require water-column correction to accurately classify zones. This eliminates the need for bathymetry data, which can be notoriously challenging to acquire.

The primary challenge with satellite images lies in their inherent variability. Images of the same site captured at different times exhibit variations in reflectance values. Although preprocessing algorithms help mitigate this effect, complete elimination remains elusive. To address this, we advocate training the models with multiple images, as demonstrated with the Sentinel-2 model, which was trained using data from eight distinct images.

Importantly, the workflow presented in this work is not limited to coral reefs but can be broadly applied to various domains, provided users possess the necessary training dataset (a full ground-truth map). Potential applications span land-use analysis, forestry, or ice sheet monitoring.

While this workflow is well-suited for long-term automated surveys, it should be complemented with an automated data acquisition and curation workflow. One prominent limitation users may currently encounter is the need to source cloud-free images. Simplified access to satellite imagery with varying temporal resolutions can complete the tool's capacity for monitoring the dynamic evolution of coral reefs. This enhancement would empower users to conduct comparative analyses by studying the presence maps of a specific site across multiple temporal frames.

Furthermore, the workflow as presented here generates static maps, which are essentially matrices of values. There's potential for future enhancements to develop interactive maps, making them more accessible outside of academia. A potential extension would be to use the compositional information embedded in the labels to generate maps that incorporate uncertainty. This could encompass dynamic maps or maps with varying color gradients to represent this uncertainty [257].

In conclusion, the workflow presented in this thesis is operational and adaptable, with ample room for refinement and expansion to enhance its usability and accessibility to a wider audience.

Bibliography

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua and S. Süsstrunk, 'Slic superpixels compared to state-of-the-art superpixel methods,' *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[2] G. E. Acquah, B. K. Via, O. O. Fasina, S. Adhikari, N. Billor and L. G. Eckhardt, 'Chemometric modeling of thermogravimetric data for the compositional analysis of forest biomass,' *PloS one*, vol. 12, no. 3, e0172999, 2017.

[3] A. Ahmad, U. K. M. Hashim, O. Mohd, M. M. Abdullah, H. Sakidin, A. W. Rasib and S. F. Sufahani, 'Comparative analysis of support vector machine, maximum likelihood and neural network classification on multispectral remote sensing data,' *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 9, pp. 529–537, 2018.

[4] A. F. Ahmed, F. N. Mutua and B. K. Kenduiywo, 'Monitoring benthic habitats using lyzenga model features from landsat multi-temporal images in google earth engine,' *Modeling Earth Systems and Environment*, pp. 1–7, 2020.

[5] J. Aitchison, 'The statistical analysis of compositional data,' *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 44, no. 2, pp. 139–160, 1982.

[6] J. Aitchison, C. Barceló-Vidal, J. A. Martín-Fernández and V. Pawlowsky-Glahn, 'Logratio analysis and compositional distance,' *Mathematical Geology*, vol. 32, no. 3, pp. 271–275, 2000.

[7] A. Akbari Asanjan, K. Das, A. Li, V. Chirayath, J. Torres-Perez and S. Sorooshian, 'Learning instrument invariant characteristics for generating high-resolution global coral reef maps,' in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2617–2624.

[8] M. L. M. Akhlaq and G. Winarso, 'Comparative analysis of object-based and pixel-based classification of high-resolution remote sensing images for mapping coral reef geomorphic zones,' in *1st Borobudur International Symposium on Humanities, Economics and Social Sciences (BIS-HESS 2019)*, Atlantis Press, 2020, pp. 992–996.

[9] R. Albright, L. Caldeira, J. Hosfelt, L. Kwiatkowski, J. K. Maclaren, B. M. Mason, Y. Nebuchina, A. Ninokawa, J. Pongratz, K. L. Ricke *et al.*, 'Reversal of ocean acidification enhances net coral reef calcification,' *Nature*, vol. 531, no. 7594, pp. 362–365, 2016.

[10] R. E. Almond, M. Grooten and T. Peterson, *Living Planet Report 2020-Bending the curve of biodiversity loss*. World Wildlife Fund, 2020.

[11] L. Alvarez-Filip, N. K. Dulvy, J. A. Gill, I. M. Côté and A. R. Watkinson, 'Flattening of caribbean coral reefs: Region-wide declines in architectural complexity,' *Proceedings of the Royal Society B: Biological Sciences*, vol. 276, no. 1669, pp. 3019–3025, 2009.

[12] E. E. Ampou, S. Ouillon, C. Iovan and S. Andréfouët, 'Change detection of bunaken island coral reefs using 15 years of very high resolution satellite images: A kaleidoscope of habitat trajectories,' *Marine pollution bulletin*, vol. 131, pp. 83–95, 2018.

[13] S. Andrefouet, F. E. Muller-Karger, J. A. Robinson, C. J. Kranenburg, D. Torres-Pulliza, S. A. Spraggins and B. Murch, 'Global assessment of modern coral reef extent and diversity for regional science and management applications: A view from space,' in *Proceedings of the 10th International Coral Reef Symposium*, Japanese Coral Reef Society Okinawa, Japan, vol. 2, 2006, pp. 1732–1745.

[14] S. Andréfouët and O. Bionaz, 'Lessons from a global remote sensing mapping project. a review of the impact of the millennium coral reef mapping project for science and management,' *Science of The Total Environment*, p. 145 987, 2021.

[15] S. Andréfouët, P. Mumby, M. McField, C. Hu and F. Muller-Karger, 'Revisiting coral reef connectivity,' *Coral Reefs*, vol. 21, no. 1, pp. 43–48, 2002.

[16] S. Andréfouët and O. Bionaz, 'Lessons from a global remote sensing mapping project. A review of the impact of the Millennium Coral Reef Mapping Project for science and management,' *Science of the Total Environment*, vol. 776, p. 145 987, 2021, ISSN: 18791026. DOI: 10.1016/j.scitotenv.2021.145987. [Online]. Available: https://doi.org/10.1016/j.scitotenv.2021.145987.

[17] S. Andréfouët, 'Coral reef habitat mapping using remote sensing: A user vs producer perspective. implications for research, management and capacity building,' *Journal of Spatial Science*, vol. 53, no. 1, pp. 113–129, 2008.

[18] S. Andréfouët, M. M. Guillaume, A. Delval, F. Rasoamanendrika, J. Blanchot and J. H. Bruggemann, 'Fifty years of changes in reef flat habitats of the grand récif of toliara (sw madagascar) and the impact of gleaning,' *Coral Reefs*, vol. 32, no. 3, pp. 757–768, 2013.

[19] S. Andréfouët, F. E. Muller-Karger, E. J. Hochberg, C. Hu and K. L. Carder, 'Change detection in shallow coral reef environments using landsat 7 etm+ data,' *Remote Sensing of Environment*, vol. 78, no. 1-2, pp. 150–162, 2001.

[20] A. Anggoro, E. Sumartono, V. Siregar, S. Agus, D. Purnama, D. Puspitosari, T. Listyorini, B. Sulistyo *et al.*, 'Comparing object-based and pixel-based classifications for benthic habitats mapping in pari islands,' vol. 1114, no. 1, p. 012 049, 2018.

[21] L. Anselin, 'Model validation in spatial econometrics: A review and evaluation of alternative approaches,' *International Regional Science Review*, vol. 11, no. 3, pp. 279–316, 1988.

[22] L. Anselin, *Spatial econometrics: methods and models*. Springer Science & Business Media, 1988, vol. 4.

[23] L. Anselin and A. K. Bera, 'Spatial dependence in linear regression models with an introduction to spatial econometrics,' *Statistics textbooks and monographs*, vol. 155, pp. 237–290, 1998.

[24] A. Ansper and K. Alikas, 'Retrieval of chlorophyll a from sentinel-2 msi data for the european union water framework directive reporting purposes,' *Remote Sensing*, vol. 11, no. 1, p. 64, 2019.

[25] D. Archer, H. Kheshgi and E. Maier-Reimer, 'Dynamics of fossil fuel co2 neutralization by marine caco3,' *Global Biogeochemical Cycles*, vol. 12, no. 2, pp. 259–276, 1998.

[26] A. Ariasari, P. Wicaksono *et al.*, 'Random forest classification and regression for seagrass mapping using planetscope image in labuan bajo, east nusa tenggara,' vol. 11372, 113721Q, 2019.

[27] R. B. Aronson, J. F. Bruno, W. F. Precht, P. W. Glynn, C. D. Harvell, L. Kaufman, C. S. Rogers, E. A. Shinn and J. F. Valentine, 'Causes of coral reef degradation,' *Science*, vol. 302, no. 5650, pp. 1502–1504, 2003.

[28] H. El-Askary, S. Abd El-Mawla, J. Li, M. El-Hattab and M. El-Raey, 'Change detection of coral reef habitat using landsat-5 tm, landsat 7 etm+ and landsat 8 oli data in the red sea (hurghada, egypt),' *International journal of remote sensing*, vol. 35, no. 6, pp. 2327–2346, 2014.

[29] G. P. Asner, R. E. Martin and J. Mascaro, 'Coral reef atoll assessment in the south china sea using planet dove satellites,' *Remote Sensing in Ecology and Conservation*, vol. 3, no. 2, pp. 57–65, 2017.

[30] Z. S. Aulia, T. T. Ahmad, R. R. Ayustina, F. T. Hastono, R. R. Hidayat, H. Mustakin, A. Fitrianto and F. B. Rifanditya, 'Shallow water seabed profile changes in 2016-2018 based on landsat 8 satellite imagery (case study: Semak daun island, karya island and gosong balik layar),' *Omni-Akuatika*, vol. 16, no. 3, pp. 26–32, 2020.

[31] G. B. R. M. P. Authority, *Geomorphological nomenclature: reef cover and zonation on the Great Barrier Reef*. Great Barrier Reef Marine Park Authority, 1986.

[32] M. B. Lyons, C. M. Roelfsema, E. V. Kennedy, E. M. Kovacs, R. Borrego-Acevedo, K. Markey, M. Roe, D. M. Yuwono, D. L. Harris, S. R. Phinn *et al.*, 'Mapping the world's coral reefs using a global multiscale earth observation framework,' *Remote Sensing in Ecology and Conservation*, 2020.

[33] L. Baetens, C. Desjardins and O. Hagolle, 'Validation of copernicus sentinel-2 cloud masks obtained from maja, sen2cor, and fmask processors using reference cloud masks generated with a supervised active learning procedure,' *Remote Sensing*, vol. 11, no. 4, p. 433, 2019.

[34] T. Bajjouk, P. Mouquet, M. Ropert, J.-P. Quod, L. Hoarau, L. Bigot, N. Le Dantec, C. Delacourt and J. Populus, 'Detection of changes in shallow coral reefs status: Towards a spatial approach using hyperspectral and multispectral data,' *Ecological Indicators*, vol. 96, pp. 174–191, 2019.

[35] B. H. Baltagi, *Econometric analysis of panel data*. Springer, 2008, vol. 4.

[36] A. Barendregt, T. Zeegers, W. van Steenis and E. Jongejans, 'Forest hoverfly community collapse: Abundance and species richness drop over four decades,' *Insect Conservation and Diversity*, vol. 15, no. 5, pp. 510–521, 2022.

[37] S. Barve, J. M. Webster and R. Chandra, 'Reef-insight: A framework for reef habitat mapping with clustering methods using remote sensing,' *Information*, vol. 14, no. 7, p. 373, 2023.

[38] G. E. Batista, R. C. Prati and M. C. Monard, 'A study of the behavior of several methods for balancing machine learning training data,' *ACM SIGKDD explorations newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[39] L. Beaudrot, J. A. Ahumada, T. O'Brien, P. Alvarez-Loayza, K. Boekee, A. Campos-Arceiz, D. Eichberg, S. Espinosa, E. Fegraus, C. Fletcher *et al.*, 'Standardized assessment of biodiversity trends in tropical forest protected areas: The end is not in sight,' *PLoS biology*, vol. 14, no. 1, e1002357, 2016.

[40] M. Beck, I. Losada, P. Menéndez, B. Reguero, P. Díaz-Simal and F. Fernández, *The global flood protection savings provided by coral reefs. nat commun 9: 2186*, 2018.

[41] S. Bejarano, P. J. Mumby, J. D. Hedley and I. Sotheran, 'Combining optical and acoustic data to enhance the detection of caribbean forereef habitats,' *Remote Sensing of Environment*, vol. 114, no. 11, pp. 2768–2778, 2010.

[42] M. Belgiu and L. Drăguţ, 'Random forest in remote sensing: A review of applications and future directions,' *ISPRS journal of photogrammetry and remote sensing*, vol. 114, pp. 24–31, 2016.

[43] M. Belgiu and J. Thomas, 'Ontology based interpretation of very high resolution imageries–grounding ontologies on visual interpretation keys,' *AGILE 2013—Leuven*, pp. 14–17, 2013.

[44] S. Benabdelkader and F. Melgani, 'Contextual spatiospectral postreconstruction of cloud-contaminated images,' *IEEE Geoscience and remote sensing letters*, vol. 5, no. 2, pp. 204–208, 2008.

[45] S. Benfield, H. Guzman, J. Mair and J. Young, 'Mapping the distribution of coral reefs and associated sublittoral habitats in pacific panama: A comparison of optical satellite sensors and classification methodologies,' *International Journal of Remote Sensing*, vol. 28, no. 22, pp. 5047–5070, 2007.

[46] T. M. Berhane, C. R. Lane, Q. Wu, B. C. Autrey, O. A. Anenkhonov, V. V. Chepinoga and H. Liu, 'Decision-tree, rule-based, and random forest classification of high-resolution multispectral imagery for wetland mapping and inventory,' *Remote sensing*, vol. 10, no. 4, p. 580, 2018.

[47] K. J. Beron and W. P. Vijverberg, 'Probit in a spatial context: A monte carlo analysis,' *Advances in spatial econometrics: methodology, tools and applications*, pp. 169–195, 2004.

[48] M. L. Berumen and M. S. Pratchett, 'Recovery without resilience: Persistent disturbance and long-term shifts in the structure of fish and coral communities at tiahura reef, moorea,' *Coral reefs*, vol. 25, no. 4, pp. 647–653, 2006.

[49] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader and J. Chanussot, 'Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches,' *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 354–379, 2012. DOI: 10.1109/JSTARS.2012.2194696.

[50] D. Blakeway and M. G. Hamblin, 'Self-generated morphology in lagoon reefs,' *PeerJ*, vol. 2015, no. 5, pp. 1–30, 2015, ISSN: 21678359. DOI: 10.7717/peerj.935.

[51] E. J. Botha, V. E. Brando, J. M. Anstee, A. G. Dekker and S. Sagar, 'Increased spectral resolution enhances coral detection under varying water conditions,' *Remote Sensing of Environment*, vol. 131, pp. 247–261, 2013.

[52] J. Boucher and D. Friot, *Primary microplastics in the oceans: a global evaluation of sources*. Iucn Gland, Switzerland, 2017, vol. 10.

[53] C. Bountzouklis, D. M. Fox and E. Di Bernardino, 'Predicting wildfire ignition causes in southern france using explainable artificial intelligence (xai) methods,' *Environmental Research Letters*, vol. 18, no. 4, p. 044 038, 2023.

[54] D. Breitburg, L. A. Levin, A. Oschlies, M. Grégoire, F. P. Chavez, D. J. Conley, V. Garçon, D. Gilbert, D. Gutiérrez, K. Isensee *et al.*, 'Declining oxygen in the global ocean and coastal waters,' *Science*, vol. 359, no. 6371, eaam7240, 2018.

[55] M. Brisset, S. Van Wynsberge, S. Andréfouët, C. Payri, B. Soulard, E. Bourassin, R. L. Gendre and E. Coutures, 'Hindcast and near real-time monitoring of green macroalgae blooms in shallow coral reef lagoons using sentinel-2: A new-caledonia case study,' *Remote Sensing*, vol. 13, no. 2, p. 211, 2021.

[56] J. F. Brown, H. J. Tollerud, C. P. Barber, Q. Zhou, J. L. Dwyer, J. E. Vogelmann, T. R. Loveland, C. E. Woodcock, S. V. Stehman, Z. Zhu *et al.*, 'Lessons learned implementing an operational continuous united states national land change monitoring capability: The land change monitoring, assessment, and projection (lcmap) approach,' *Remote Sensing of Environment*, vol. 238, p. 111 356, 2020.

[57] A. Bruckner, G. Rowlands, B. Riegl, S. J. Purkis, A. Williams and P. Renaud, 'Atlas of saudi arabian red sea marine habitats,' 2013.

[58] A. Buccianti and V. Pawlowsky-Glahn, 'New perspectives on water chemistry and compositional data analysis,' *Mathematical Geology*, vol. 37, pp. 703–727, 2005.

[59] B. Bulgarelli, V. Kiselev and G. Zibordi, 'Adjacency effects in satellite radiometric products from coastal waters: A theoretical analysis for the northern adriatic sea,' *Applied Optics*, vol. 56, no. 4, pp. 854–869, 2017.

[60] C. Bunkhumpornpat, K. Sinapiromsaran and C. Lursinsap, 'Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem,' in *Pacific-Asia conference on knowledge discovery and data mining*, Springer, 2009, pp. 475–482.

[61] H. K. Burgess, L. DeBey, H. Froehlich, N. Schmidt, E. J. Theobald, A. K. Ettinger, J. HilleRisLambers, J. Tewksbury and J. K. Parrish, 'The science of citizen science: Exploring barriers to use as a primary research tool,' *Biological Conservation*, vol. 208, pp. 113–120, 2017.

[62] B. Burkhard, F. Kroll, S. Nedkov and F. Müller, 'Mapping ecosystem service supply, demand and budgets,' *Ecological indicators*, vol. 21, pp. 17–29, 2012.

[63] C. Burns, B. Bollard and A. Narayanan, 'Machine-learning for mapping and monitoring shallow coral reef habitats,' *Remote Sensing*, vol. 14, no. 11, p. 2666, 2022.

[64] J. H. Burns, D. Delparte, R. D. Gates and M. Takabayashi, 'Integrating structure-from-motion photogrammetry with geospatial software as a novel technique for quantifying 3D ecological characteristics of coral reefs,' *PeerJ*, vol. 2015, no. 7, e1077, 2015, ISSN: 21678359. DOI: 10.7717/peerj.1077.

[65] J. Busch, L. Greer, D. Harbor, K. Wirth, H. Lescinsky, H. A. Curran and K. de Beurs, 'Quantifying exceptionally large populations of acropora spp. corals off belize using sub-meter satellite imagery classification,' *Bulletin of Marine Science*, vol. 92, no. 2, p. 265, 2016.

[66] J. D. Butler, S. J. Purkis, R. Yousif, I. Al-Shaikh and C. Warren, 'A high-resolution remotely sensed benthic habitat map of the qatari coastal zone,' *Marine Pollution Bulletin*, vol. 160, p. 111 634, 2020.

[67] G. Cabioch, B. Thomassin and J. Lecolle, 'Age d'émersion des récifs frangeants holocènes autour de la Grande Terre de Nouvelle-Calédonie (SO Pacifique). Nouvelle interprétation de la courbe des niveaux marins depuis 8000 ans B.P,' *Comptes Rendus de l'Académie des Sciences.Série 2 : Mécanique...*, vol. 308, no. 4, pp. 419–425, 1989, ISSN: 0249-6305.

[68] R. Calabrese and J. A. Elkink, 'Estimators of binary spatial autoregressive models: A monte carlo study,' *Journal of Regional Science*, vol. 54, no. 4, pp. 664–687, 2014.

[69] L. Camacho, G. Douzas and F. Bacao, 'Geometric smote for regression,' *Expert Systems with Applications*, vol. 193, p. 116 387, 2022.

[70] G. Cao, P. C. Kyriakidis and M. F. Goodchild, 'A multinomial logistic mixed model for the prediction of categorical spatial data,' *International Journal of Geographical Information Science*, vol. 25, no. 12, pp. 2071–2086, 2011.

[71] T. Caras, J. Hedley and A. Karnieli, 'Implications of sensor design for coral reef detection: Upscaling ground hyperspectral imagery in spatial and spectral scales,' *International journal of applied earth observation and geoinformation*, vol. 63, pp. 68–77, 2017.

[72] J. Carlot, M. Vousdoukas, A. Rovere, T. Karambas, H. S. Lenihan, M. Kayal, M. Adjeroud, G. Pérez-Rosales, L. Hedouin and V. Parravicini, 'Coral reef structural complexity loss exposes coastlines to waves,' *Scientific Reports*, vol. 13, no. 1, p. 1683, 2023.

[73] M. Chami, X. Lenot, M. Guillaume, B. Lafrance, X. Briottet, A. Minghelli, S. Jay, Y. Deville and V. Serfaty, 'Analysis and quantification of seabed adjacency effects in the subsurface upward radiance in shallow waters,' *Optics express*, vol. 27, no. 8, A319–A338, 2019.

[74] F. Charte, A. J. Rivera, M. J. del Jesus and F. Herrera, 'Mlsmote: Approaching imbalanced multilabel learning through synthetic instance generation,' *Knowledge-Based Systems*, vol. 89, pp. 385–397, 2015.

[75] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, 'Smote: Synthetic minority over-sampling technique,' *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[76] A. Chegoonian, M. Mokhtarzade and M. Valadan Zoej, 'A comprehensive evaluation of classification algorithms for coral reef habitat mapping: Challenges related to quantity, quality, and impurity of training samples,' *International Journal of Remote Sensing*, vol. 38, no. 14, pp. 4224–4243, 2017.

[77] A. M. Chegoonian, M. Mokhtarzade, M. J. V. Zoej and M. Salehi, 'Soft supervised classification: An improved method for coral reef classification using medium resolution satellite images,' pp. 2787–2790, 2016.

[78] A. Chemchem, F. Alin and M. Krajecki, 'Combining smote sampling and machine learning for forecasting wheat yields in france,' in *2019 IEEE second international conference on artificial intelligence and knowledge engineering (AIKE)*, IEEE, 2019, pp. 9–14.

[79] A. Chen, Y. Ma and J. Zhang, 'Partition satellite derived bathymetry for coral reefs based on spatial residual information,' *International Journal of Remote Sensing*, vol. 42, no. 8, pp. 2807–2826, 2021.

[80] X. Chen, L. Vierling and D. Deering, 'A simple and effective radiometric correction method to improve landscape change detection across sensors and across time,' *Remote Sensing of Environment*, vol. 98, no. 1, pp. 63–79, 2005.

[81] A. Chin, ''hunting porcupines': Citizen scientists contribute new knowledge about rare coral reef species.,' *Pacific Conservation Biology*, vol. 20, no. 1, pp. 48–53, 2014.

[82] V. Chirayath, 'Nemo-net & fluid lensing: The neural multi-modal observation & training network for global coral reef assessment using fluid lensing augmentation of nasa eos data,' 2018.

[83] V. Chirayath and A. Li, 'Next-generation optical sensing technologies for exploring ocean worlds—nasa fluidcam, midar, and nemo-net,' *Frontiers in Marine Science*, vol. 6, p. 521, 2019.

[84] J. D. Clare, P. A. Townsend, C. Anhalt-Depies, C. Locke, J. L. Stenglein, S. Frett, K. J. Martin, A. Singh, T. R. Van Deelen and B. Zuckerberg, 'Making inference with messy (citizen science) data: When are data accurate enough and how can they be improved?' *Ecological Applications*, vol. 29, no. 2, e01849, 2019.

[85] A. D. Cliff and K. Ord, 'Spatial autocorrelation: A review of existing and new measures with applications,' *Economic Geography*, vol. 46, no. sup1, pp. 269–292, 1970.

[86] I. P. on Climate Change, *Climate change: The IPCC response strategies*. World Meteorological Organization, 1990.

[87] J. P. Coakley and B. Rust, 'Sedimentation in an arctic lake,' *Journal of Sedimentary Research*, vol. 38, no. 4, pp. 1290–1300, 1968.

[88] M. M. Coffer, B. A. Schaeffer, R. C. Zimmerman, V. Hill, J. Li, K. A. Islam and P. J. Whitman, 'Performance across worldview-2 and rapideye for reproducible seagrass mapping,' *Remote Sensing of Environment*, vol. 250, p. 112 036, 2020.

[89] A. Collin, M. Andel, D. James and J. Claudet, 'The superspectral/hyperspatial worldview-3 as the link between spaceborn hyperspectral and airborne hyperspatial sensors: The case study of the complex tropical coast,' *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2019.

[90] A. Collin, M. Andel, D. Lecchini and J. Claudet, 'Mapping sub-metre 3d land-sea coral reefscapes using superspectral worldview-3 satellite stereoimagery,' in *Oceans*, Multidisciplinary Digital Publishing Institute, vol. 2, 2021, pp. 315–329.

[91] A. Collin, P. Archambault and S. Planes, 'Bridging ridge-to-reef patches: Seamless classification of the coast using very high resolution satellite,' *Remote Sensing*, vol. 5, no. 7, pp. 3583–3610, 2013.

[92] A. Collin, J. L. Hench, Y. Pastol, S. Planes, L. Thiault, R. J. Schmitt, S. J. Holbrook, N. Davies and M. Troyer, 'High resolution topobathymetry using a pleiades-1 triplet: Moorea island in 3d,' *Remote sensing of environment*, vol. 208, pp. 109–119, 2018.

[93] A. Collin, J. Laporte, B. Koetz, F.-R. Martin-Lauzer and Y.-L. Desnos, 'Coral reefs in fatu huku island, marquesas archipelago, french polynesia,' in *Seafloor Geomorphology as Benthic Habitat*, Elsevier, 2020, pp. 533–543.

[94] A. Collin, C. Ramambason, Y. Pastol, E. Casella, A. Rovere, L. Thiault, B. Espiau, G. Siu, F. Lerouvreur, N. Nakamura *et al.*, 'Very high resolution mapping of coral reef state using airborne bathymetric lidar surface-intensity and drone imagery,' *International journal of remote sensing*, vol. 39, no. 17, pp. 5676–5688, 2018.

[95] L. A. Conti, G. T. da Mota and R. L. Barcellos, 'High-resolution optical remote sensing for coastal benthic habitat mapping: A case study of the suape estuarine-bay, pernambuco, brazil,' *Ocean & Coastal Management*, vol. 193, p. 105 205, 2020.

[96] M. J. Costello, A. Cheung and N. De Hauwere, 'Surface area and the seabed area, volume, depth, slope, and topographic variation for the world's seas, oceans, and countries,' *Environmental science & technology*, vol. 44, no. 23, pp. 8821–8828, 2010.

[97] M. Crowson, E. Warren-Thomas, J. K. Hill, B. Hariyadi, F. Agus, A. Saad, K. C. Hamer, J. A. Hodgson, W. D. Kartika, J. Lucey *et al.*, 'A comparison of satellite remote sensing data fusion methods to map peat swamp forest loss in sumatra, indonesia,' *Remote Sensing in Ecology and Conservation*, vol. 5, no. 3, pp. 247–258, 2019.

[98] T. Cui, J. Zhang, K. Wang, J. Wei, B. Mu, Y. Ma, J. Zhu, R. Liu and X. Chen, 'Remote sensing of chlorophyll a concentration in turbid coastal waters based on a global optical water classification system,' *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 163, pp. 187–201, 2020.

[99] M. Dalponte, H. O. Ørka, T. Gobakken, D. Gianelle and E. Næsset, 'Tree species classification in boreal forests with hyperspectral data,' *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 5, pp. 2632–2645, 2012.

[100] R. Danovaro, E. Fanelli, J. Aguzzi, D. Billett, L. Carugati, C. Corinaldesi, A. Dell'Anno, K. Gjerde, A. J. Jamieson, S. Kark *et al.*, 'Ecological variables for developing a global deep-ocean monitoring and conservation strategy,' *Nature Ecology & Evolution*, vol. 4, no. 2, pp. 181–192, 2020.

[101] C. Darwin, *Geological observations on coral reefs, volcanic islands, and on South America: Being the geology of the voyage of the Beagle, under the command of Captain Fitzroy, RN, during the years 1832 to 1836*. Smith, Elder & Company, 1851.

[102] K. A. Davis, G. Pawlak and S. G. Monismith, 'Turbulence and coral reefs,' *Annual review of marine science*, vol. 13, pp. 343–373, 2021.

[103] G. De'ath, J. M. Lough and K. E. Fabricius, 'Declining coral calcification on the great barrier reef,' *Science*, vol. 323, no. 5910, pp. 116–119, 2009.

[104] M. Deng, Y. Guo, C. Wang and F. Wu, 'An oversampling method for multi-class imbalanced data based on composite weights,' *PloS one*, vol. 16, no. 11, e0259227, 2021.

[105] M. DeSalvo, A. Estrada, S. Sunagawa and M. Medina, 'Transcriptomic responses to darkness stress point to common coral bleaching mechanisms,' *Coral Reefs*, vol. 31, no. 1, pp. 215–228, 2012.

[106] C. F. Dormann, J. M. McPherson, M. B. Araújo, R. Bivand, J. Bolliger, G. Carl, R. G. Davies, A. Hirzel, W. Jetz, W. D. Kissling *et al.*, 'Methods to account for spatial autocorrelation in the analysis of species distributional data: A review,' *Ecography*, pp. 609–628, 2007.

[107] S. Duce, R. J. Mccarroll, B. Yiu, L. A. Perris, B. Yumba, S. Wales and E. Melbourne, 'Field measurements from contrasting reefs show spurs and grooves can dissipate more wave energy than the reef crest,' *Earth and Space Science Open Archive ESSOAr*, 2021.

[108] S. Duce, A. Vila-Concejo, S. M. Hamylton, J. M. Webster, E. Bruce and R. J. Beaman, 'A morphometric assessment and classification of coral reef spur and groove morphology,' *Geomorphology*, vol. 265, pp. 68–83, 2016, ISSN: 0169555X. DOI: 10.1016/j.geomorph.2016.04.018.

[109] N. K. Dulvy, N. Pacoureau, C. L. Rigby, R. A. Pollom, R. W. Jabado, D. A. Ebert, B. Finucci, C. M. Pollock, J. Cheok, D. H. Derrick *et al.*, 'Overfishing drives over one-third of all sharks and rays toward a global extinction crisis,' *Current Biology*, vol. 31, no. 21, pp. 4773–4787, 2021.

[110] S. Durant, T. Wacher, S. Bashir, R. d. Woodroffe, P. De Ornellas, C. Ransom, J. Newby, T. Abáigar, M. Abdelgadir, H. El Alqamy *et al.*, 'Fiddling in biodiversity hotspots while deserts burn? collapse of the s ahara's megafauna,' *Diversity and Distributions*, vol. 20, no. 1, pp. 114–122, 2014.

[111] P. Ebel, A. Meraner, M. Schmitt and X. X. Zhu, 'Multisensor data fusion for cloud removal in global and all-season sentinel-2 imagery,' *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[112] F. Eugenio, J. Marcello and J. Martin, 'High-resolution maps of bathymetry and benthic habitats in shallow-water environments using multispectral remote sensing imagery,' *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3539–3549, 2015.

[113] F. Eugenio, J. Marcello, J. Martin and D. Rodríguez-Esparragón, 'Benthic habitat mapping using multispectral high-resolution imagery: Evaluation of shallow water atmospheric correction techniques,' *Sensors*, vol. 17, no. 11, p. 2639, 2017.

[114] F. Eugenio, J. Marcello, A. Mederos-Barrera and F. Marqués, 'High-resolution satellite bathymetry mapping: Regression and machine learning-based approaches,' *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[115] A. R. Fahlevi, T. Osawa and I. W. Arthana, 'Coral reef and shallow water benthic identification using landsat 7 etm+ satellite data in nusa penida district,' *International Journal of Environment and Geosciences*, vol. 2, no. 1, pp. 17–34, 2018.

[116] F. R. de Faria, D. Barbosa, C. A. Howe, K. L. R. Canabrava, J. E. Sasaki and P. R. dos Santos Amorim, 'Time-use movement behaviors are associated with scores of depression/anxiety among adolescents: A compositional data analysis,' *Plos one*, vol. 17, no. 12, e0279401, 2022.

[117] P. F. Felzenszwalb and D. P. Huttenlocher, 'Efficient graph-based image segmentation,' *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[118] L. Feng and C. Hu, 'Cloud adjacency effects on top-of-atmosphere radiance and ocean color data products: A statistical assessment,' *Remote Sensing of Environment*, vol. 174, pp. 301–313, 2016.

[119] A. Fernández, S. Garcia, F. Herrera and N. V. Chawla, 'Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary,' *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, 2018.

[120] S. A. Foo and G. P. Asner, 'Impacts of remotely sensed environmental drivers on coral outplant survival,' *Restoration Ecology*, vol. 29, no. 1, e13309, 2021.

[121] G. Forkuor, C. Conrad, M. Thiel, T. Landmann and B. Barry, 'Evaluating the sequential masking classification approach for improving crop discrimination in the sudanian savanna of west africa,' *Computers and Electronics in Agriculture*, vol. 118, pp. 380–389, 2015.

[122] G. Forrester, P. Baily, D. Conetta, L. Forrester, E. Kintzing and L. Jarecki, 'Comparing monitoring data collected by volunteers and professionals shows that citizen scientists can detect long-term change on coral reefs,' *Journal for Nature Conservation*, vol. 24, pp. 1–9, 2015.

[123] B. Fox-Kemper, 'Ocean, cryosphere and sea level change,' in *AGU Fall Meeting Abstracts*, vol. 2021, 2021, U13B–09.

[124] I. Francis and J. Newton, 'Determining wine aroma from compositional data,' *Australian Journal of Grape and Wine Research*, vol. 11, no. 2, pp. 114–126, 2005.

[125] D. Frantz, E. Haß, A. Uhl, J. Stoffels and J. Hill, 'Improvement of the fmask algorithm for sentinel-2 images: Separating clouds from bright surfaces based on parallax effects,' *Remote sensing of environment*, vol. 215, pp. 471–481, 2018.

[126] R. S. Fraser and Y. J. Kaufman, 'The relative importance of aerosol scattering and absorption in remote sensing,' *IEEE transactions on geoscience and remote sensing*, no. 5, pp. 625–633, 1985.

[127] R. Froese, H. Winker, G. Coro, N. Demirel, A. C. Tsikliras, D. Dimarchopoulou, G. Scarcella, M. Quaas and N. Matz-Lück, 'Status and rebuilding of european fisheries,' *Marine Policy*, vol. 93, pp. 159–170, 2018.

[128] B. Fu, Y. Wang, A. Campbell, Y. Li, B. Zhang, S. Yin, Z. Xing and X. Jin, 'Comparison of object-based and pixel-based random forest algorithm for wetland vegetation mapping using high spatial resolution gf-1 and sar data,' *Ecological indicators*, vol. 73, pp. 105–117, 2017.

[129] S. Gabarda and G. Cristobal, 'Cloud covering denoising through image fusion,' *Image and Vision Computing*, vol. 25, no. 5, pp. 523–530, 2007.

[130] M. Galetti, C. Brocardo, R. Begotti, L. Hortenci, F. Rocha-Mendes, C. Bernardo, R. Bueno, R. Nobre, R. Bovendorp, R. Marques *et al.*, 'Defaunation and biomass collapse of mammals in the largest atlantic forest remnant,' *Animal Conservation*, vol. 20, no. 3, pp. 270–281, 2017.

[131] C. Gang, W. Zhou, Y. Chen, Z. Wang, Z. Sun, J. Li, J. Qi and I. Odeh, 'Quantitative assessment of the contributions of climate change and human activities on global grassland degradation,' *Environmental Earth Sciences*, vol. 72, pp. 4273–4282, 2014.

[132] J. J. Gapper, H. El-Askary, E. Linstead and T. Piechota, 'Evaluation of spatial generalization characteristics of a robust classifier as applied to coral reef habitats in remote islands of the pacific ocean,' *Remote Sensing*, vol. 10, no. 11, p. 1774, 2018.

[133] J. J. Gapper, H. El-Askary, E. Linstead and T. Piechota, 'Coral reef change detection in remote pacific islands using support vector machine classifiers,' *Remote Sensing*, vol. 11, no. 13, p. 1525, 2019.

[134] R. A. Garcia, Z. Lee and E. J. Hochberg, 'Hyperspectral shallow-water remote sensing with an enhanced benthic classifier,' *Remote Sensing*, vol. 10, no. 1, p. 147, 2018.

[135] J.-P. Gattuso, A. K. Magnan, L. Bopp, W. W. Cheung, C. M. Duarte, J. Hinkel, E. Mcleod, F. Micheli, A. Oschlies, P. Williamson *et al.*, 'Ocean solutions to address climate change and its effects on marine ecosystems,' *Frontiers in Marine Science*, p. 337, 2018.

[136] M. Y. Gazi, T. J. Mowsumi and M. K. Ahmed, 'Detection of coral reefs degradation using geospatial techniques around saint martin's island, bay of bengal,' *Ocean Science Journal*, vol. 55, no. 3, pp. 419–431, 2020.

[137] S. Georganos, T. Grippa, A. Niang Gadiaga, C. Linard, M. Lennert, S. Vanhuysse, N. Mboga, E. Wolff and S. Kalogirou, 'Geographical random forests: A spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling,' *Geocarto International*, vol. 36, no. 2, pp. 121–136, 2021.

[138] M. Gholoum, D. Bruce and S. Alhazeem, 'A new image classification approach for mapping coral density in state of kuwait using high spatial resolution satellite images,' *International Journal of Remote Sensing*, vol. 40, no. 12, pp. 4787–4816, 2019.

[139] A. Ghosh and P. K. Joshi, 'A comparison of selected classification algorithms for mapping bamboo patches in lower gangetic plains using very high resolution worldview 2 imagery,' *International Journal of Applied Earth Observation and Geoinformation*, vol. 26, pp. 298–311, 2014.

[140] M. Gianinetto and M. Scaioni, 'Automated geometric correction of high-resolution pushbroom satellite data,' *Photogrammetric Engineering & Remote Sensing*, vol. 74, no. 1, pp. 107–116, 2008.

[141] C. Giardino, M. Bresciani, F. Fava, E. Matta, V. E. Brando and R. Colombo, 'Mapping submerged habitats and mangroves of lampi island marine national park (myanmar) from in situ and satellite observations,' *Remote Sensing*, vol. 8, no. 1, p. 2, 2016.

[142] R. Gibson, R. Atkinson, J. Gordon, I. Smith and D. Hughes, 'Coral-associated invertebrates: Diversity, ecological importance and vulnerability to disturbance,' *Oceanogr. Mar. Biol*, vol. 49, pp. 43–104, 2011.

[143] E. Gischler, 'Indo-Pacific and Atlantic spurs and grooves revisited: The possible effects of different Holocene sea-level history, exposure, and reef accretion rate in the shallow fore reef,' *Facies*, vol. 56, no. 2, pp. 173–177, 2010, ISSN: 01729179. DOI: 10.1007/s10347-010-0218-0.

[144] M. A. Gleyzes, L. Perret and P. Kubik, 'Pleiades system architecture and main performances,' *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 39, pp. 537–542, 2012.

[145] P. Glynn and L. D'croz, 'Experimental evidence for high temperature stress as the cause of el nino-coincident coral mortality,' *Coral reefs*, vol. 8, no. 4, pp. 181–191, 1990.

[146] P. Glynn, E. C. Peters and L. Muscatine, 'Coral tissue microstructure and necrosis: Relation to catastrophic coral mortality in panama,' *Dis Aquat Org*, vol. 1, pp. 29–37, 1985.

[147] D. Gomes, A. S. Saif and D. Nandi, 'Robust underwater object detection with autonomous underwater vehicle: A comprehensive study,' in *Proceedings of the International Conference on Computing Advancements*, 2020, pp. 1–10.

[148] A. M. Gomez, K. C. McDonald, K. Shein, S. DeVries, R. A. Armstrong, W. J. Hernandez and M. Carlo, 'Comparison of satellite-based sea surface temperature to in situ observations surrounding coral reefs in la parguera, puerto rico,' *Journal of Marine Science and Engineering*, vol. 8, no. 6, p. 453, 2020.

[149] P. C. Gonzalez-Espinosa and S. D. Donner, 'Cloudiness reduces the bleaching response of coral reefs exposed to heat stress,' *Global Change Biology*, vol. 27, no. 15, pp. 3474–3486, 2021.

[150] M. Gonzalez-Rivero, C. Roelfsema, S. Lopez-Marcano, C. Castro-Sanguino, T. Bridge and R. Babcock, 'Supplementary report to the final report of the coral reef expert group: S6. novel technologies in coral reef monitoring,' 2020.

[151] M. González-Rivero, O. Beijbom, A. Rodriguez-Ramirez, T. Holtrop, Y. González-Marrero, A. Ganase, C. Roelfsema, S. Phinn and O. Hoegh-Guldberg, 'Scaling up ecological measurements of coral reefs using semi-automated field image collection and analysis,' *Remote Sensing*, vol. 8, no. 1, p. 30, 2016.

[152] C. Gonzalo-Martin, A. Garcia-Pedrero, M. Lillo-Saavedra and E. Menasalvas, 'Deep learning for superpixel-based classification of remote sensing images,' 2016.

[153] J. A. Goodman, M. Lay, L. Ramirez, S. L. Ustin and P. J. Haverkamp, 'Confidence levels, sensitivity, and the role of bathymetry in coral reef remote sensing,' *Remote Sensing*, vol. 12, no. 3, p. 496, 2020.

[154] J. A. Goodman, S. J. Purkis and S. R. Phinn, 'Coral reef remote sensing,' *A guide for mapping, monitoring and management. 436p*, 2013.

[155] H. R. Gordon, 'Removal of atmospheric effects from satellite imagery of the oceans,' *Applied Optics*, vol. 17, no. 10, pp. 1631–1636, 1978.

[156] H. R. Gordon, 'Atmospheric correction of ocean color imagery in the earth observing system era,' *Journal of Geophysical Research: Atmospheres*, vol. 102, no. D14, pp. 17 081–17 106, 1997.

[157] H. R. Gordon and M. Wang, 'Surface-roughness considerations for atmospheric correction of ocean color sensors. 1: The rayleigh-scattering component,' *Applied optics*, vol. 31, no. 21, pp. 4247–4260, 1992.

[158] H. R. Gordon and M. Wang, 'Retrieval of water-leaving radiance and aerosol optical thickness over the oceans with seawifs: A preliminary algorithm,' *Applied optics*, vol. 33, no. 3, pp. 443–452, 1994.

[159] E. Gordon-Rodriguez, T. Quinn and J. P. Cunningham, 'Data augmentation for compositional data: Advancing predictive models of the microbiome,' *Advances in Neural Information Processing Systems*, vol. 35, pp. 20 551–20 565, 2022.

[160] M. Goulard, T. Laurent and C. Thomas-Agnan, 'About predictions in spatial autoregressive models: Optimal and almost optimal strategies,' *Spatial Economic Analysis*, vol. 12, no. 2-3, pp. 304–325, 2017.

[161] N. Graham and K. Nash, 'The importance of structural complexity in coral reef ecosystems,' *Coral reefs*, vol. 32, no. 2, pp. 315–326, 2013.

[162] P. C. Gray, J. T. Ridge, S. K. Poulin, A. C. Seymour, A. M. Schwantes, J. J. Swenson and D. W. Johnston, 'Integrating drone imagery into high resolution satellite remote sensing assessments of estuarine environments,' *Remote Sensing*, vol. 10, no. 8, p. 1257, 2018.

[163] G. Groom, R. Fuller and A. Jones, 'Contextual correction: Techniques for improving land cover mapping from remotely sensed images,' *International Journal of Remote Sensing*, vol. 17, no. 1, pp. 69–89, 1996.

[164] M. Guillaume, Y. Michels and S. Jay, 'Joint estimation of water column parameters and seabed reflectance combining maximum likelihood and unmixing algorithm,' *2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pp. 1–4, 2015.

[165] M. Guillaume, A. Minghelli, Y. Deville, M. Chami, L. Juste, X. Lenot, B. Lafrance, S. Jay, X. Briottet and V. Serfaty, 'Mapping benthic habitats by extending non-negative matrix factorization to address the water column and seabed adjacency effects,' *Remote Sensing*, vol. 12, no. 13, p. 2072, 2020.

[166] N. T. Ha, M. Manley-Harris, T. D. Pham and I. Hawes, 'A comparative assessment of ensemble-based machine learning and maximum likelihood methods for mapping seagrass using sentinel-2 imagery in tauranga harbor, new zealand,' *Remote Sensing*, vol. 12, no. 3, p. 355, 2020.

[167] C. M. Hafner, 'The spread of the covid-19 pandemic in time and space,' *International journal of environmental research and public health*, vol. 17, no. 11, p. 3827, 2020.

[168] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue and G. Bing, 'Learning from class-imbalanced data: Review of methods and applications,' *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017.

[169] C. A. Hallmann, M. Sorg, E. Jongejans, H. Siepel, N. Hofland, H. Schwan, W. Stenmans, A. Müller, H. Sumser, T. Hörren *et al.*, 'More than 75 percent decline over 27 years in total flying insect biomass in protected areas,' *PloS one*, vol. 12, no. 10, e0185809, 2017.

[170] S. M. Hamylton, S. Duce, A. Vila-Concejo, C. M. Roelfsema, S. R. Phinn, R. C. Carvalho, E. C. Shaw and K. E. Joyce, 'Estimating regional coral reef calcium carbonate production from remotely sensed seafloor maps,' *Remote sensing of environment*, vol. 201, pp. 88–98, 2017.

[171] S. Hamylton and J. Mallela, 'Reef development on a remote coral atoll before and after coral bleaching: A geospatial assessment,' *Marine Geology*, vol. 418, p. 106 041, 2019.

[172] H. Han, W.-Y. Wang and B.-H. Mao, 'Borderline-smote: A new over-sampling method in imbalanced data sets learning,' in *International conference on intelligent computing*, Springer, 2005, pp. 878–887.

[173] J. Hand and W. Malm, 'Review of aerosol mass scattering efficiencies from ground-based measurements since 1990,' *Journal of Geophysical Research: Atmospheres*, vol. 112, no. D16, 2007.

[174] M. C. Hansen and T. R. Loveland, 'A review of large area monitoring of land cover change using landsat data,' *Remote sensing of Environment*, vol. 122, pp. 66–74, 2012.

[175] P. T. Harris and E. K. Baker, 'Why map benthic habitats?' In *Seafloor geomorphology as benthic habitat*, Elsevier, 2012, pp. 3–22.

[176] J. Hedley, A. Harborne and P. Mumby, 'Simple and robust removal of sun glint for mapping shallow-water benthos,' *International Journal of Remote Sensing*, vol. 26, no. 10, pp. 2107–2112, 2005.

[177] J. D. Hedley, C. Roelfsema, V. Brando, C. Giardino, T. Kutser, S. Phinn, P. J. Mumby, O. Barrilero, J. Laporte and B. Koetz, 'Coral reef applications of sentinel-2: Coverage, characteristics, bathymetry and benthic mapping with comparison to landsat 8,' *Remote sensing of environment*, vol. 216, pp. 598–614, 2018.

[178] J. D. Hedley, C. M. Roelfsema, I. Chollett, A. R. Harborne, S. F. Heron, S. J. Weeks, W. J. Skirving, A. E. Strong, C. M. Eakin, T. R. Christensen *et al.*, 'Remote sensing of coral reefs for monitoring and management: A review,' *Remote Sensing*, vol. 8, no. 2, p. 118, 2016.

[179] J. D. Hedley, C. M. Roelfsema, S. R. Phinn and P. J. Mumby, 'Environmental and sensor limitations in optical remote sensing of coral reefs: Implications for monitoring and sensor design,' *Remote Sensing*, vol. 4, no. 1, pp. 271–302, 2012.

[180] M. Helmi, A. Aysira, M. Munasik, A. Wirasatriya, R. Widiaratih and R. Ario, 'Spatial structure analysis of benthic ecosystem based on geospatial approach at parang islands, karimunjawa national park, central java, indonesia,' *Indonesian Journal of Oceanography*, vol. 2, no. 1, pp. 40–47, 2020.

[181] S. S. Heydari and G. Mountrakis, 'Effect of classifier selection, reference sample size, reference class distribution and scene heterogeneity in per-pixel classification accuracy using 26 landsat sites,' *Remote Sensing of Environment*, vol. 204, pp. 648–658, 2018.

[182] J.-P. A. Hobbs, Z. T. Richards, I. Popovic, C. Lei, T. M. Staeudle, S. R. Montanari and J. D. DiBattista, 'Hybridisation and the evolution of coral reef biodiversity,' *Coral Reefs*, vol. 41, no. 3, pp. 535–549, 2022.

[183] O. Hoegh-Guldberg, P. J. Mumby, A. J. Hooten, R. S. Steneck, P. Greenfield, E. Gomez, C. D. Harvell, P. F. Sale, A. J. Edwards, K. Caldeira *et al.*, 'Coral reefs under rapid climate change and ocean acidification,' *science*, vol. 318, no. 5857, pp. 1737–1742, 2007.

[184] R. K. Hoeke, C. D. Storlazzi and P. V. Ridd, 'Drivers of circulation in a fringing coral reef embayment: A wave-flow coupled numerical modeling study of Hanalei Bay, Hawaii,' *Continental Shelf Research*, vol. 58, pp. 79–95, 2013, ISSN: 02784343. DOI: 10.1016/j.csr.2013.03.007.

[185] K. T. Holland, M. L. Palmsten *et al.*, 'Remote sensing applications and bathymetric mapping in coastal environments,' *Advances in Coastal Hydraulics*, pp. 375–411, 2018.

[186] J. Holloway-Brown, K. J. Helmstedt and K. L. Mengersen, 'Stochastic spatial random forest (ss-rf) for interpolating probabilities of missing land cover data,' *Journal of Big Data*, vol. 7, no. 1, pp. 1–23, 2020.

[187] J. Holloway-Brown, K. J. Helmstedt and K. L. Mengersen, 'Spatial random forest (s-rf): A random forest approach for spatially interpolating missing land-cover data with multiple classes,' *International Journal of Remote Sensing*, vol. 42, no. 10, pp. 3756–3776, 2021.

[188] P. Holthus and J. E. Maragos, 'Marine ecosystem classification for the tropical island pacific,' *Marine and coastal biodiversity in the tropical island Pacific region*, vol. 1, pp. 239–278, 1995.

[189] M. S. Hossain, A. M. Muslim, M. I. Nadzri, A. W. Sabri, I. Khalil, Z. Mohamad and A. Beiranvand Pour, 'Coral habitat mapping: A comparison between maximum likelihood, bayesian and dempster–shafer classifiers,' *Geocarto International*, pp. 1–19, 2019.

[190] M. S. Hossain, A. M. Muslim, M. I. Nadzri, K. Teruhisa, D. David, I. Khalil and Z. Mohamad, 'Can ensemble techniques improve coral reef habitat classification accuracy using multispectral data?' *Geocarto International*, vol. 35, no. 11, pp. 1214–1232, 2020.

[191] P. Houk and R. van Woesik, 'Dynamics of shallow-water assemblages in the saipan lagoon,' *Marine Ecology Progress Series*, vol. 356, pp. 39–50, 2008.

[192] Z. X. Hoy, K. S. Woon, W. C. Chin, H. Hashim and Y. Van Fan, 'Forecasting heterogeneous municipal solid waste generation via bayesian-optimised neural network with ensemble learning for improved generalisation,' *Computers & Chemical Engineering*, vol. 166, p. 107 946, 2022.

[193] C. Huang, N. Thomas, S. N. Goward, J. G. Masek, Z. Zhu, J. R. Townshend and J. E. Vogelmann, 'Automated masking of cloud and cloud shadow for forest change analysis using landsat images,' *International Journal of Remote Sensing*, vol. 31, no. 20, pp. 5449–5464, 2010.

[194] Y. Huang, D.-R. Liu, S.-J. Lee, C.-H. Hsu and Y.-G. Liu, 'A boosting resampling method for regression based on a conditional variational autoencoder,' *Information Sciences*, vol. 590, pp. 90–105, 2022.

[195] M. J. Hughes and D. J. Hayes, 'Automated detection of cloud and cloud shadow in single-date landsat imagery using neural networks and spatial post-processing,' *Remote Sensing*, vol. 6, no. 6, pp. 4907–4926, 2014.

[196] T. P. Hughes, K. D. Anderson, S. R. Connolly, S. F. Heron, J. T. Kerry, J. M. Lough, A. H. Baird, J. K. Baum, M. L. Berumen, T. C. Bridge *et al.*, 'Spatial and temporal patterns of mass bleaching of corals in the anthropocene,' *Science*, vol. 359, no. 6371, pp. 80–83, 2018.

[197] T. P. Hughes, J. T. Kerry, A. H. Baird, S. R. Connolly, T. J. Chase, A. Dietzel, T. Hill, A. S. Hoey, M. O. Hoogenboom, M. Jacobson *et al.*, 'Global warming impairs stock–recruitment dynamics of corals,' *Nature*, vol. 568, no. 7752, pp. 387–390, 2019.

[198] T. P. Hughes, J. T. Kerry, A. H. Baird, S. R. Connolly, A. Dietzel, C. M. Eakin, S. F. Heron, A. S. Hoey, M. O. Hoogenboom, G. Liu *et al.*, 'Global warming transforms coral reef assemblages,' *Nature*, vol. 556, no. 7702, pp. 492–496, 2018.

[199] D. Ierodiaconou, A. C. Schimel, D. Kennedy, J. Monk, G. Gaylard, M. Young, M. Diesing and A. Rattray, 'Combining pixel and object based image analysis of ultra-high resolution multibeam bathymetry and backscatter for habitat mapping in shallow marine waters,' *Marine Geophysical Research*, vol. 39, pp. 271–288, 2018.

[200] K. J. Iknayan and S. R. Beissinger, 'Collapse of a desert bird community over the past century driven by climate change,' *Proceedings of the National Academy of Sciences*, vol. 115, no. 34, pp. 8597–8602, 2018.

[201] M. Immitzer, C. Atzberger and T. Koukal, 'Tree species classification with random forest using very high spatial resolution 8-band worldview-2 satellite data,' *Remote sensing*, vol. 4, no. 9, pp. 2661–2693, 2012.

[202] F. Immordino, M. Barsanti, E. Candigliota, S. Cocito, I. Delbono and A. Peirano, 'Application of sentinel-2 multispectral data for habitat mapping of pacific islands: Palau republic (micronesia, pacific ocean),' *Journal of Marine Science and Engineering*, vol. 7, no. 9, p. 316, 2019.

[203] W. IPBES, 'Intergovernmental science-policy platform on biodiversity and ecosystem services,' *Summary for Policy Makers of the Global Assessment Report on Biodiversity and Ecosystem Services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. IPBES Secretariat, Bonn, Germany*, 2019.

[204] A. Iqbal, W. A. Qazi, N. Shahzad and M. Nazeer, 'Identification and mapping of coral reefs using landsat 8 oli in astola island, pakistan coastal ocean,' pp. 1–6, 2018.

[205] D. A. Jackson, 'Compositional data in community ecology: The paradigm or peril of proportions?' *Ecology*, vol. 78, no. 3, pp. 929–940, 1997.

[206] J. R. Jambeck, R. Geyer, C. Wilcox, T. R. Siegler, M. Perryman, A. Andrady, R. Narayan and K. L. Law, 'Plastic waste inputs from land into the ocean,' *Science*, vol. 347, no. 6223, pp. 768–771, 2015.

[207] J. Jarrett, I. Saleh, M. B. Blake, R. Malcolm, S. Thorpe and T. Grandison, 'Combining human and machine computing elements for analysis via crowdsourcing,' in *10th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing*, IEEE, 2014, pp. 312–321.

[208] J. H. Jeppesen, R. H. Jacobsen, F. Inceoglu and T. S. Toftegaard, 'A cloud detection algorithm for satellite imagery based on deep learning,' *Remote sensing of environment*, vol. 229, pp. 247–259, 2019.

[209] J. M. Johnson and T. M. Khoshgoftaar, 'Survey on deep learning with class imbalance,' *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[210] M. I. Jordan and T. M. Mitchell, 'Machine learning: Trends, perspectives, and prospects,' *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[211] J. Ju and D. P. Roy, 'The availability of cloud-free landsat etm+ data over the conterminous united states and globally,' *Remote Sensing of Environment*, vol. 112, no. 3, pp. 1196–1211, 2008.

[212] K. Kabiri, H. Rezai and M. Moradi, 'Mapping of the corals around hendorabi island (persian gulf), using worldview-2 standard imagery coupled with field observations,' *Marine pollution bulletin*, vol. 129, no. 1, pp. 266–274, 2018.

[213] P. S. Kanaroglou, M. D. Adams, P. F. De Luca, D. Corr and N. Sohel, 'Estimation of sulfur dioxide air pollution concentrations with a spatial autoregressive model,' *Atmospheric Environment*, vol. 79, pp. 421–427, 2013.

[214] D. Kaur and Y. Kaur, 'Various image segmentation techniques: A review,' *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 5, pp. 809–814, 2014.

[215] S. Kay, J. D. Hedley and S. Lavender, 'Sun glint correction of high and low spatial resolution images of aquatic scenes: A review of methods for visible and near-infrared wavelengths,' *Remote sensing*, vol. 1, no. 4, pp. 697–730, 2009.

[216] E. V. Kennedy, C. M. Roelfsema, M. B. Lyons *et al.*, 'Reef Cover, a coral reef classification for global habitat mapping from remote sensing,' *Scientific Data*, vol. 8, no. 1, pp. 1–20, 2021. DOI: 10.1038/s41597-021-00958-z.

[217] R. E. Kennedy, S. Andréfouët, W. B. Cohen, C. Gómez, P. Griffiths, M. Hais, S. P. Healey, E. H. Helmer, P. Hostert, M. B. Lyons *et al.*, 'Bringing an ecological view of change to landsat-based remote sensing,' *Frontiers in Ecology and the Environment*, vol. 12, no. 6, pp. 339–346, 2014.

[218] K. Kerrigan and K. A. Ali, 'Application of landsat 8 oli for monitoring the coastal waters of the us virgin islands,' *International Journal of Remote Sensing*, vol. 41, no. 15, pp. 5743–5769, 2020.

[219] M. H. Kesikoglu, U. H. Atasever, F. Dadaser-Celik and C. Ozkan, 'Performance of ann, svm and mlh techniques for land use/cover change detection at sultan marshes wetland, turkey,' *Water Science and Technology*, vol. 80, no. 3, pp. 466–477, 2019.

[220] R. Khatami, G. Mountrakis and S. V. Stehman, 'A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research,' *Remote Sensing of Environment*, vol. 177, pp. 89–100, 2016.

[221] J. A. Kleypas and K. K. Yates, 'Coral reefs and ocean acidification,' *Oceanography*, vol. 22, no. 4, pp. 108–117, 2009.

[222] T. Kogut, A. Tomczak, A. Słowik and T. Oberski, 'Seabed modelling by means of airborne laser bathymetry data and imbalanced learning for offshore mapping,' *Sensors*, vol. 22, no. 9, p. 3121, 2022.

[223] A. Kollert, M. Bremer, M. Löw and M. Rutzinger, 'Exploring the potential of land surface phenology and seasonal cloud free composites of one year of sentinel-2 imagery for tree species mapping in a mountainous region,' *International Journal of Applied Earth Observation and Geoinformation*, vol. 94, p. 102 208, 2021.

[224] P. K. Koner, A. Harris and E. Maturi, 'Hybrid cloud and error masking to improve the quality of deterministic satellite sea surface temperature retrieval and data coverage,' *Remote Sensing of Environment*, vol. 174, pp. 266–278, 2016.

[225] N. A. Kornder, B. M. Riegl and J. Figueiredo, 'Thresholds and drivers of coral calcification responses to climate change,' *Global change biology*, vol. 24, no. 11, pp. 5084–5095, 2018.

[226] E. Kovacs, C. Roelfsema, M. Lyons, S. Zhao and S. Phinn, 'Seagrass habitat mapping: How do landsat 8 oli, sentinel-2, zy-3a, and worldview-3 perform?' *Remote Sensing Letters*, vol. 9, no. 7, pp. 686–695, 2018.

[227] T. Krisztin, P. Piribauer and M. Wögerer, 'A spatial multinomial logit model for analysing urban expansion,' *Spatial Economic Analysis*, vol. 17, no. 2, pp. 223–244, 2022.

[228] K. J. Kroeker, R. L. Kordas, R. Crim, I. E. Hendriks, L. Ramajo, G. S. Singh, C. M. Duarte and J.-P. Gattuso, 'Impacts of ocean acidification on marine organisms: Quantifying sensitivities and interaction with warming,' *Global change biology*, vol. 19, no. 6, pp. 1884–1896, 2013.

[229] P. Kumar, D. K. Gupta, V. N. Mishra and R. Prasad, 'Comparison of support vector machine, artificial neural network, and spectral angle mapper algorithms for crop classification using liss iv data,' *International Journal of Remote Sensing*, vol. 36, no. 6, pp. 1604–1617, 2015.

[230] T. Kutser, B. Paavel, K. Kaljurand, M. Ligi and M. Randla, 'Mapping shallow waters of the baltic sea with sentinel-2 imagery,' pp. 1–6, 2018.

[231] C. Kwan, B. Budavari, A. C. Bovik and G. Marchisio, 'Blind quality assessment of fused worldview-3 images by using the combinations of pansharpening and hypersharpening paradigms,' *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1835–1839, 2017.

[232] H. S. Ladd, 'Recent Reefs,' *AAPG Bulletin*, vol. 34, no. 2, pp. 203–214, 1950, ISSN: 0149-1423. DOI: 10.1306/3d933ecf-16b1-11d7-8645000102c1865d.

[233] J. B. Lamb, B. L. Willis, E. A. Fiorenza, C. S. Couch, R. Howard, D. N. Rader, J. D. True, L. A. Kelly, A. Ahmad, J. Jompa *et al.*, 'Plastic waste associated with disease on coral reefs,' *Science*, vol. 359, no. 6374, pp. 460–462, 2018.

[234] W. F. Laurance, D. Carolina Useche, J. Rendeiro, M. Kalka, C. J. Bradshaw, S. P. Sloan, S. G. Laurance, M. Campbell, K. Abernethy, P. Alvarez *et al.*, 'Averting biodiversity collapse in tropical forest protected areas,' *Nature*, vol. 489, no. 7415, pp. 290–294, 2012.

[235] W. Lazuardi, P. Wicaksono and M. Marfai, 'Remote sensing for coral reef and seagrass cover mapping to support coastal management of small islands,' in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing, vol. 686, 2021, p. 012 031.

[236] T. J. Leininger, A. E. Gelfand, J. M. Allen and J. A. Silander, 'Spatial regression modeling for compositional data with many zeros,' *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 18, pp. 314–334, 2013.

[237] M. P. Lesser and J. H. Farrell, 'Exposure to solar radiation increases damage to both host tissues and algal symbionts of corals during thermal stress,' *Coral reefs*, vol. 23, no. 3, pp. 367–377, 2004.

[238] A. S. Levine and C. L. Feinholz, 'Participatory gis to inform coral reef ecosystem management: Mapping human coastal and ocean uses in hawaii,' *Applied Geography*, vol. 59, pp. 60–69, 2015.

[239] A. S. Li, V. Chirayath, M. Segal-Rozenhaimer, J. L. Torres-Perez and J. van den Bergh, 'Nasa nemo-net's convolutional neural network: Mapping marine habitats with spectrally heterogeneous remote sensing imagery,' *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5115–5133, 2020.

[240] J. Li, N. S. Fabina, D. E. Knapp and G. P. Asner, 'The sensitivity of multi-spectral satellite sensors to benthic habitat change,' *Remote Sensing*, vol. 12, no. 3, p. 532, 2020.

[241] J. Li, D. E. Knapp, N. S. Fabina, E. V. Kennedy, K. Larsen, M. B. Lyons, N. J. Murray, S. R. Phinn, C. M. Roelfsema and G. P. Asner, 'A global coral reef probability map generated using convolutional neural networks,' *Coral Reefs*, vol. 39, pp. 1805–1815, 2020.

[242] J. Li, S. R. Schill, D. E. Knapp and G. P. Asner, 'Object-based mapping of coral reef habitats using planet dove satellites,' *Remote Sensing*, vol. 11, no. 12, p. 1445, 2019.

[243] M. Li, J. Wu and X. Deng, 'Identifying drivers of land use change in china: A spatial multinomial logit model analysis,' *Land Economics*, vol. 89, no. 4, pp. 632–654, 2013.

[244] X.-M. Li, Y. Ma, Z.-H. Leng, J. Zhang and X.-X. Lu, 'High-accuracy remote sensing water depth retrieval for coral islands and reefs based on lstm neural network,' *Journal of Coastal Research*, vol. 102, no. SI, pp. 21–32, 2020.

[245] Z.-L. Li, H. Wu, N. Wang, S. Qiu, J. A. Sobrino, Z. Wan, B.-H. Tang and G. Yan, 'Land surface emissivity retrieval from satellite data,' *International Journal of Remote Sensing*, vol. 34, no. 9-10, pp. 3084–3127, 2013.

[246] A. Lim, A. J. Wheeler and L. Conti, 'Cold-water coral habitat mapping: Trends and developments in acquisition and processing methods,' *Geosciences*, vol. 11, no. 1, p. 9, 2021.

[247] R. A. de Lima, A. A. Oliveira, G. R. Pitta, A. L. de Gasper, A. C. Vibrans, J. Chave, H. Ter Steege and P. I. Prado, 'The erosion of biodiversity and biomass in the atlantic forest biodiversity hotspot,' *Nature communications*, vol. 11, no. 1, p. 6347, 2020.

[248] C.-H. Lin, P.-H. Tsai, K.-H. Lai and J.-Y. Chen, 'Cloud removal from multitemporal satellite images using information cloning,' *IEEE transactions on geoscience and remote sensing*, vol. 51, no. 1, pp. 232–241, 2012.

[249] C. Lin, C.-C. Wu, K. Tsogt, Y.-C. Ouyang and C.-I. Chang, 'Effects of atmospheric correction and pansharpening on lulc classification accuracy using worldview-2 imagery,' *Information Processing in Agriculture*, vol. 2, no. 1, pp. 25–36, 2015.

[250] G. Liu, C. M. Eakin, M. Chen, A. Kumar, J. L. De La Cour, S. F. Heron, E. F. Geiger, W. J. Skirving, K. V. Tirak and A. E. Strong, 'Predicting heat stress to inform reef management: Noaa coral reef watch's 4-month coral bleaching outlook,' *Frontiers in Marine Science*, vol. 5, p. 57, 2018.

[251] Y. Liu, P. Song, J. Peng and C. Ye, 'A physical explanation of the variation in threshold for delineating terrestrial water surfaces from multi-temporal images: Effects of radiometric correction,' *International Journal of Remote Sensing*, vol. 33, no. 18, pp. 5862–5875, 2012.

[252] J. L. Loerzel, T. L. Goedeke, M. K. Dillard and G. Brown, 'Scuba divers above the waterline: Using participatory mapping of coral reef conditions to inform reef management,' *Marine Policy*, vol. 76, pp. 79–89, 2017.

[253] C. A. Logan, J. P. Dunne, C. M. Eakin and S. D. Donner, 'Incorporating adaptive responses into future projections of coral bleaching,' *Global Change Biology*, vol. 20, no. 1, pp. 125–139, 2014.

[254] S. Long, B. Sparrow-Scinocca, M. E. Blicher, N. Hammeken Arboe, M. Fuhrmann, K. M. Kemp, R. Nygaard, K. Zinglersen and C. Yesson, 'Identification of a soft coral garden candidate vulnerable marine ecosystem (vme) using video imagery, davis strait, west greenland,' *Frontiers in Marine Science*, vol. 7, p. 460, 2020.

[255] G. M. Lovett, D. A. Burns, C. T. Driscoll, J. C. Jenkins, M. J. Mitchell, L. Rustad, J. B. Shanley, G. E. Likens and R. Haeuber, 'Who needs environmental monitoring?' *Frontiers in Ecology and the Environment*, vol. 5, no. 5, pp. 253–260, 2007.

[256] R. J. Lowe, J. L. Falter, S. G. Monismith and M. J. Atkinson, 'Wave-driven circulation of a coastal reef-lagoon system,' *Journal of Physical Oceanography*, vol. 39, no. 4, pp. 873–893, 2009, ISSN: 00223670. DOI: 10.1175/2008JPO3958.1.

[257] L. R. Lucchesi, P. M. Kuhnert and C. K. Wikle, 'Vizumap: An r package for visualising uncertainty in spatial data,' *Journal of Open Source Software*, vol. 6, no. 59, p. 2409, 2021.

[258] D. R. Lyzenga, 'Passive remote sensing techniques for mapping water depth and bottom features,' *Applied optics*, vol. 17, no. 3, pp. 379–383, 1978.

[259] D. R. Lyzenga, N. P. Malinas and F. J. Tanis, 'Multispectral bathymetry using a simple physically based algorithm,' *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 8, pp. 2251–2259, 2006.

[260] S. Ma, C. Zhou, C. Chi, Y. Liu and G. Yang, 'Estimating physical composition of municipal solid waste in china by applying artificial neural network method,' *Environmental science & technology*, vol. 54, no. 15, pp. 9609–9617, 2020.

[261] P. Maglione, C. Parente and A. Vallario, 'Coastline extraction using high resolution worldview-2 satellite imagery,' *European Journal of Remote Sensing*, vol. 47, no. 1, pp. 685–699, 2014.

[262] R. Magno, L. Rocchi, R. Dainelli, A. Matese, S. F. Di Gennaro, C.-F. Chen, N.-T. Son and P. Toscano, 'Agroshadow: A new sentinel-2 cloud shadow detection tool for precision agriculture,' *Remote Sensing*, vol. 13, no. 6, p. 1219, 2021.

[263] M. Mahdianpari, B. Salehi, F. Mohammadimanesh and M. Motagh, 'Random forest wetland classification using alos-2 l-band, radarsat-2 c-band, and terrasar-x imagery,' *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 130, pp. 13–31, 2017.

[264] A. Mahmood, M. Bennamoun, S. An, F. Sohel and F. Boussaid, 'Resfeats: Residual network based features for underwater image classification,' *Image and Vision Computing*, vol. 93, p. 103 811, 2020.

[265] M. Maier, 'Dirichletreg: Dirichlet regression for compositional data in r,' 2014.

[266] M. D. M. Manessa, A. Kanno, M. Sekine, E. E. Ampou, N. Widagti, A. As-syakur *et al.*, 'Shallow-water benthic identification using multispectral satellite imagery: Investigation on the effects of improving noise correction method and spectral cover,' *Remote Sensing*, vol. 6, no. 5, pp. 4454–4472, 2014.

[267] J. Marcello, F. Eugenio and F. Marqués, 'Benthic mapping using high resolution multispectral and hyperspectral imagery,' in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 2018, pp. 1535–1538.

[268] G. Marre, C. D. A. Braga, D. Ienco, S. Luque, F. Holon and J. Deter, 'Deep convolutional neural networks to monitor coralligenous reefs: Operationalizing biodiversity and ecological assessment,' *Ecological Informatics*, vol. 59, p. 101 110, 2020.

[269] N. J. Marshall, D. A. Kleine and A. J. Dean, 'Coralwatch: Education, monitoring, and sustainability through citizen science,' *Frontiers in Ecology and the Environment*, vol. 10, no. 6, pp. 332–334, 2012.

[270] P. Marshall and A. Baird, 'Bleaching of corals on the great barrier reef: Differential susceptibilities among taxa,' *Coral reefs*, vol. 19, no. 2, pp. 155–163, 2000.

[271] P. Martimort, O. Arino, M. Berger, R. Biasutti, B. Carnicero, U. Del Bello, V. Fernandez, F. Gascon, B. Greco, P. Silvestrin *et al.*, 'Sentinel-2 optical high resolution mission for gmes operational services,' in *2007 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 2007, pp. 2677–2680.

[272] J. Martin, F. Eugenio, J. Marcello and A. Medina, 'Automatic sun glint removal of multispectral high-resolution worldview-2 imagery for retrieving coastal shallow water parameters,' *Remote Sensing*, vol. 8, no. 1, p. 37, 2016.

[273] M. L. Martínez, A. Intralawan, G. Vázquez, O. Pérez-Maqueo, P. Sutton and R. Landgrave, 'The coasts of our world: Ecological, economic and social importance,' *Ecological economics*, vol. 63, no. 2-3, pp. 254–272, 2007.

[274] F. S. Marzano, M. Iacobelli, M. Orlandi and D. Cimini, 'Coastal water remote sensing from sentinel-2 satellite data using physical, statistical, and neural network retrieval approach,' *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[275] K. McIntyre, K. McLaren and K. Prospere, 'Mapping shallow nearshore benthic features in a caribbean marine-protected area: Assessing the efficacy of using different data types (hydroacoustic versus satellite images) and classification techniques,' *International Journal of Remote Sensing*, vol. 39, no. 4, pp. 1117–1150, 2018.

[276] K. McLaren, K. McIntyre and K. Prospere, 'Using the random forest algorithm to integrate hydroacoustic data with satellite images to improve the mapping of shallow nearshore benthic features in a marine protected area in jamaica,' *GIScience & Remote Sensing*, vol. 56, no. 7, pp. 1065–1092, 2019.

[277] K. Mengersen, E. E. Peterson, S. Clifford, N. Ye, J. Kim, T. Bednarz, R. Brown, A. James, J. Vercelloni, A. R. Pearse *et al.*, 'Modelling imperfect presence data obtained by citizen science,' *Environmetrics*, vol. 28, no. 5, e2446, 2017.

[278] A. M. Mielke, 'Using deep convolutional neural networks to classify littoral areas with 3-band and 5-band imagery,' Naval Postgraduate School Monterey United States, Tech. Rep., 2020.

[279] A. Minghelli, J. Spagnoli, M. Lei, M. Chami and S. Charmasson, 'Shoreline extraction from worldview2 satellite data in the presence of foam pixels using multispectral classification method,' *Remote Sensing*, vol. 12, no. 16, p. 2664, 2020.

[280] K. Mizuno, K. Terayama, S. Hagino, S. Tabeta, S. Sakamoto, T. Ogawa, K. Sugimoto and H. Fukami, 'An efficient coral survey method based on a large-scale 3-d structure model obtained by speedy sea scanner and u-net segmentation,' *Scientific reports*, vol. 10, no. 1, pp. 1–11, 2020.

[281] M. Modasshir and I. Rekleitis, 'Enhancing coral reef monitoring utilizing a deep semi-supervised learning approach,' in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2020, pp. 1874–1880.

[282] M. A. Mograne, C. Jamet, H. Loisel, V. Vantrepotte, X. Mériaux and A. Cauvin, 'Evaluation of five atmospheric correction algorithms over french optically-complex waters for the sentinel-3a olci ocean color sensor,' *Remote Sensing*, vol. 11, no. 6, p. 668, 2019.

[283] H. Mohamed, K. Nadaoka and T. Nakamura, 'Assessment of machine learning algorithms for automatic benthic cover monitoring and mapping using towed underwater video camera and high-resolution satellite images,' *Remote Sensing*, vol. 10, no. 5, p. 773, 2018.

[284] H. Mohamed, K. Nadaoka and T. Nakamura, 'Towards benthic habitat 3d mapping using machine learning algorithms and structures from motion photogrammetry,' *Remote Sensing*, vol. 12, no. 1, p. 127, 2020.

[285] N. R. Mollica, W. Guo, A. L. Cohen, K.-F. Huang, G. L. Foster, H. K. Donald and A. R. Solow, 'Ocean acidification affects coral growth by reducing skeletal density,' *Proceedings of the National Academy of Sciences*, vol. 115, no. 8, pp. 1754–1759, 2018.

[286] N. Moniz, R. Ribeiro, V. Cerqueira and N. Chawla, 'Smoteboost for regression: Improving the prediction of extreme values,' in *2018 IEEE 5th international conference on data science and advanced analytics (DSAA)*, IEEE, 2018, pp. 150–159.

[287] E. F. Morales and H. J. Escalante, 'A brief introduction to supervised, unsupervised, and reinforcement learning,' in *Biosignal processing and classification using computational learning and intelligence*, Elsevier, 2022, pp. 111–129.

[288] F. E. Muller-Karger, E. Hestir, C. Ade, K. Turpie, D. A. Roberts, D. Siegel, R. J. Miller, D. Humm, N. Izenberg, M. Keller *et al.*, 'Satellite sensor requirements for monitoring essential biodiversity variables of coastal ecosystems,' *Ecological applications*, vol. 28, no. 3, pp. 749–760, 2018.

[289] P. J. Mumby and A. J. Edwards, 'Mapping marine environments with ikonos imagery: Enhanced spatial resolution can deliver greater thematic accuracy,' *Remote sensing of environment*, vol. 82, no. 2-3, pp. 248–257, 2002.

[290] P. Mumby, C. Clark, E. Green and A. J. Edwards, 'Benefits of water column correction and contextual editing for mapping coral reefs,' *international Journal of Remote sensing*, vol. 19, no. 1, pp. 203–210, 1998.

[291] A. M. Muslim, W. S. Chong, C. D. M. Safuan, I. Khalil and M. S. Hossain, 'Coral reef mapping of uav: A comparison of sun glint correction methods,' *Remote Sensing*, vol. 11, no. 20, p. 2422, 2019.

[292] R. Naidu, F. Muller-Karger and M. McCarthy, 'Mapping of benthic habitats in komave, coral coast using worldview-2 satellite imagery,' in *Climate Change Impacts and Adaptation Strategies for Coastal Communities*, Springer, 2018, pp. 337–355.

[293] M. A. Najar, G. Thoumyre, E. W. J. Bergsma, R. Almar, R. Benshila and D. G. Wilson, 'Satellite derived bathymetry using deep learning,' *Machine Learning*, pp. 1–24, 2021.

[294] N. Nandakumar, T. Baldwin and B. Salehi, 'How well do embedding models capture non-compositionality? a view from multiword expressions,' in *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, 2019, pp. 27–34.

[295] A. P. Negri, F. Flores, T. Röthig and S. Uthicke, 'Herbicides increase the vulnerability of corals to rising sea surface temperature,' *Limnology and Oceanography*, vol. 56, no. 2, pp. 471–485, 2011.

[296] P. Neubert and P. Protzel, 'Compact watershed and preemptive slic: On improving trade-offs of superpixel segmentation algorithms,' in *2014 22nd international conference on pattern recognition*, IEEE, 2014, pp. 996–1001.

[297] L. H. Nguyen, D. R. Joshi, D. E. Clay and G. M. Henebry, 'Characterizing land cover/land use from multiple years of landsat and modis time series: A novel approach using land surface phenology modeling and random forest classifier,' *Remote Sensing of Environment*, vol. 238, p. 111 017, 2020.

[298] T. Nguyen, B. Liquet, K. Mengersen and D. Sous, 'Mapping of coral reefs with multispectral satellites: A review of recent papers,' *Remote Sensing*, vol. 13, no. 21, p. 4470, 2021.

[299] T. Nguyen, K. Mengersen, D. Sous and B. Liquet, 'Smote-cd: Smote for compositional data,' *PloS one*, vol. 18, no. 6, e0287705, 2023.

[300] T. H. A. Nguyen, T. Laurent, C. Thomas-Agnan and A. Ruiz-Gazen, 'Analyzing the impacts of socio-economic factors on french departmental elections with coda methods,' *Journal of Applied Statistics*, vol. 49, no. 5, pp. 1235–1251, 2022.

[301] T. H. A. Nguyen, C. Thomas-Agnan, T. Laurent and A. Ruiz-Gazen, 'A simultaneous spatial autoregressive model for compositional data,' *Spatial Economic Analysis*, vol. 16, no. 2, pp. 161–175, 2021.

[302] S. Nichol, Z. Huang, F. Howard, R. Porter-Smith, V. Lucieer and N. Barrett, 'Geomorphological classification of reefs,' *Report to the National Environmental Science Program, Marine Biodiversity Hub*, p. 27, 2016.

[303] M. Niroumand-Jadidi and A. Vitti, 'Optimal band ratio analysis of worldview-3 imagery for bathymetry of shallow rivers (case study: Sarca river, italy),' *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci*, pp. 361–365, 2016.

[304] N. Nurdin, T. Komatsu, M. A. AS, A. R. Djalil, K. Amri *et al.*, 'Multisensor and multitemporal data from landsat images to detect damage to coral reefs, small islands in the spermonde archipelago, indonesia,' *Ocean Science Journal*, vol. 50, no. 2, pp. 317–325, 2015.

[305] M. Nurlidiasari and S. Budiman, 'Mapping coral reef habitat with and without water column correction using quickbird image,' *International Journal of Remote Sensing and Earth Sciences (IJReSES)*, vol. 2, 2010.

[306] D. O'sullivan and D. Unwin, *Geographic information analysis*. John Wiley & Sons, 2003.

[307] Y. Oktorini, V. Darlis, N. Wahidin and R. Jhonnerie, 'The use of spot 6 and rapideye imageries for mangrove mapping in the kembung river, bengkalis island, indonesia,' in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing, vol. 695, 2021, p. 012 009.

[308] J. Olafsson, S. Olafsdottir, A. Benoit-Cattin, M. Danielsen, T. Arnarson and T. Takahashi, 'Rate of iceland sea acidification from time series measurements.,' *Biogeosciences*, vol. 6, no. 11, 2009.

[309] J. C. Orr, V. J. Fabry, O. Aumont, L. Bopp, S. C. Doney, R. A. Feely, A. Gnanadesikan, N. Gruber, A. Ishida, F. Joos *et al.*, 'Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms,' *Nature*, vol. 437, no. 7059, pp. 681–686, 2005.

[310] J.-C. Ortiz, N. H. Wolff, K. R. Anthony, M. Devlin, S. Lewis and P. J. Mumby, 'Impaired recovery of the great barrier reef under cumulative stress,' *Science advances*, vol. 4, no. 7, eaar6127, 2018.

[311] A. Ozdarici-Ok, A. O. Ok and K. Schindler, 'Mapping of agricultural crops from single high-resolution multispectral images—data-driven smoothing vs. parcel-based smoothing,' *Remote Sensing*, vol. 7, no. 5, pp. 5611–5638, 2015.

[312] R. K. Pace and R. Barry, 'Fast spatial estimation,' *Applied Economics Letters*, vol. 4, no. 5, pp. 337–341, 1997.

[313] J.-C. Padró, F.-J. Muñoz, L. Á. Ávila, L. Pesquer and X. Pons, 'Radiometric correction of landsat-8 and sentinel-2a scenes using drone imagery in synergy with field spectroradiometry,' *Remote Sensing*, vol. 10, no. 11, p. 1687, 2018.

[314] D. A. Palandro, S. Andréfouët, C. Hu, P. Hallock, F. E. Müller-Karger, P. Dustan, M. K. Callahan, C. Kranenburg and C. R. Beaver, 'Quantification of two decades of shallow-water coral reef habitat decline in the florida keys national marine sanctuary using landsat data (1984–2002),' *Remote Sensing of Environment*, vol. 112, no. 8, pp. 3388–3399, 2008.

[315] L. Paolini, F. Grings, J. A. Sobrino, J. C. Jiménez Muñoz and H. Karszenbaum, 'Radiometric correction effects in landsat multi-date/multi-sensor change detection studies,' *International Journal of Remote Sensing*, vol. 27, no. 4, pp. 685–704, 2006.

[316] C. Parente and M. Pepe, 'Bathymetry from worldview-3 satellite data using radiometric band ratio,' *Acta Polytechnica*, vol. 58, no. 2, pp. 109–117, 2018.

[317] M. A. Paul, P. A. J. Rani and J. L. Manopriya, 'Gradient based aura feature extraction for coral reef classification,' *Wireless Personal Communications*, vol. 114, no. 1, pp. 149–166, 2020.

[318] D. Pauly, R. Watson and J. Alder, 'Global trends in world fisheries: Impacts on marine ecosystems and food security,' *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1453, pp. 5–12, 2005.

[319] M. Perez-Ortiz, P. A. Gutierrez, C. Hervas-Martinez and X. Yao, 'Graph-based approaches for over-sampling in the context of ordinal regression,' *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 5, pp. 1233–1245, 2014.

[320] T. Phanomsophon, N. Jaisue, A. Worphet, N. Tawinteung, B. Shrestha, J. Posom, L. Khurnpoon and P. Sirisomboon, 'Rapid measurement of classification levels of primary macronutrients in durian (durio zibethinus murray cv. mon thong) leaves using ft-nir spectrometer and comparing the effect of imbalanced and balanced data for modelling,' *Measurement*, vol. 203, p. 111 975, 2022.

[321] S. R. Phinn, C. M. Roelfsema and P. J. Mumby, 'Multi-scale, object-based image analysis for mapping geomorphic and ecological zones on coral reefs,' *International Journal of Remote Sensing*, vol. 33, no. 12, pp. 3768–3797, 2012.

[322] D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama and M. Ranagalage, 'Sentinel-2 data for land cover/use mapping: A review,' *Remote Sensing*, vol. 12, no. 14, p. 2291, 2020.

[323] P. A. Pirazzoli, 'Marine notches.,' *Sea-level research*, pp. 361–400, 1986. DOI: 10.1007/978-94-009-4215-8_12.

[324] B. Pirzamanbein, J. Lindström, A. Poska and M.-J. Gaillard, 'Modelling spatial compositional data: Reconstructions of past land cover and uncertainties,' *Spatial statistics*, vol. 24, pp. 14–31, 2018.

[325] C. Pisapia, D. Burn and M. Pratchett, 'Changes in the population and community structure of corals during recent disturbances (february 2016-october 2017) on maldivian coral reefs,' *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[326] S. J. Pittman and K. A. Brown, 'Multi-scale approach for predicting fish species distributions across coral reef seascapes,' *PLoS ONE*, vol. 6, no. 5, e20583, 2011, ISSN: 19326203. DOI: 10.1371/journal.pone.0020583.

[327] D. Poursanidis, D. Traganos, N. Chrysoulakis and P. Reinartz, 'Cubesats allow high spatiotemporal estimates of satellite-derived bathymetry,' *Remote Sensing*, vol. 11, no. 11, p. 1299, 2019.

[328] D. Poursanidis, D. Traganos, L. Teixeira, A. Shapiro and L. Muaves, 'Cloud-native seascape mapping of mozambique's quirimbas national park with sentinel-2,' *Remote Sensing in Ecology and Conservation*, 2020.

[329] D. Poursanidis, D. Traganos, L. Teixeira, A. Shapiro and L. Muaves, 'Cloud-native seascape mapping of mozambique's quirimbas national park with sentinel-2,' *Remote Sensing in Ecology and Conservation*, vol. 7, no. 2, pp. 275–291, 2021.

[330] E. Prangel, L. Kasari-Toussaint, L. Neuenkamp, N. Noreika, R. Karise, R. Marja, N. Ingerpuu, T. Kupper, L. Keerberg, E. Oja *et al.*, 'Afforestation and abandonment of semi-natural grasslands lead to biodiversity loss and a decline in ecosystem services and functions,' *Journal of Applied Ecology*, vol. 60, no. 5, pp. 825–836, 2023.

[331] M. S. Pratchett, P. Munday, S. K. Wilson, N. Graham, J. E. Cinner, D. R. Bellwood, G. P. Jones, N. Polunin and T. McClanahan, 'Effects of climate-induced coral bleaching on coral-reef fishes,' *Ecological and economic consequences. Oceanography and Marine Biology: Annual Review*, vol. 46, pp. 251–296, 2008.

[332] P. Probst, M. N. Wright and A.-L. Boulesteix, 'Hyperparameters and tuning strategies for random forest,' *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, vol. 9, no. 3, e1301, 2019.

[333] R. Pu and S. Bell, 'Mapping seagrass coverage and spatial patterns with high spatial resolution ikonos imagery,' *International Journal of Applied Earth Observation and Geoinformation*, vol. 54, pp. 145–158, 2017.

[334] E. G. Purdy, E. Gischler and A. J. Lomando, 'The Belize margin revisited. 2. Origin of Holocene antecedent topography,' *International Journal of Earth Sciences*, vol. 92, no. 4, pp. 552–572, 2003, ISSN: 14373254. DOI: 10.1007/s00531-003-0325-z.

[335] S. J. Purkis and K. E. Kohler, 'The role of topography in promoting fractal patchiness in a carbonate shelf landscape,' *Coral Reefs*, vol. 27, no. 4, pp. 977–989, 2008, ISSN: 07224028. DOI: 10.1007/s00338-008-0404-5.

[336] S. J. Purkis, G. P. Rowlands and B. M. Riegl, 'The paradox of tropical karst morphology in the coral reefs of the arid Middle East: Reply,' *Geology*, vol. 38, no. 10, pp. 227–230, 2010, ISSN: 00917613. DOI: 10.1130/G31454Y.1.

[337] S. J. Purkis, 'Remote sensing tropical coral reefs: The view from above,' *Annual review of marine science*, vol. 10, pp. 149–168, 2018.

[338] S. J. Purkis, A. C. Gleason, C. R. Purkis, A. C. Dempsey, P. G. Renaud, M. Faisal, S. Saul and J. M. Kerr, 'High-resolution habitat and bathymetry maps for 65,000 sq. km of earth's remotest coral reefs,' *Coral Reefs*, vol. 38, no. 3, pp. 467–488, 2019.

[339] R. Putra, M. Suhana, D. Kurniawn, M. Abrar, R. Siringoringo, N. Sari, H. Irawan, E. Prayetno, T. Apriadi and A. Suryanti, 'Detection of reef scale thermal stress with aqua and terra modis satellite for coral bleaching phenomena,' in *AIP Conference Proceedings*, AIP Publishing LLC, vol. 2094, 2019, p. 020 024.

[340] S. Qiu, Z. Zhu and B. He, 'Fmask 4.0: Improved cloud and cloud shadow detection in landsats 4–8 and sentinel-2 imagery,' *Remote Sensing of Environment*, vol. 231, p. 111 205, 2019.

[341] L. Qu, Z. Chen, M. Li, J. Zhi, H. Wang *et al.*, 'Accuracy improvements to pixel-based and object-based lulc classification with auxiliary datasets from google earth engine,' *Remote Sensing*, vol. 13, no. 3, p. 453, 2021.

[342] R. Rajeesh and G. Dwarakish, 'Satellite oceanography–a review,' *Aquatic Procedia*, vol. 4, pp. 165–172, 2015.

[343] E. C. Rankey, S. L. Reeder and J. R. Garza-Pérez, 'Controls on links between geomorphical and surface sedimentological variability: Aitutaki and maupiti atolls, South Pacific Ocean,' *Journal of Sedimentary Research*, vol. 81, no. 12, pp. 885–890, 2011, ISSN: 15271404. DOI: 10.2110/jsr.2011.73.

[344] A. Raphael, Z. Dubinsky, D. Iluz, J. I. Benichou and N. S. Netanyahu, 'Deep neural network recognition of shallow water corals in the gulf of eilat (aqaba),' *Scientific reports*, vol. 10, no. 1, pp. 1–11, 2020.

[345] C. E. Raymundo, M. C. Oliveira, T. d. A. Eleuterio, S. R. André, M. G. da Silva, E. R. d. S. Queiroz and R. d. A. Medronho, 'Spatial analysis of covid-19 incidence and the sociodemographic context in brazil,' *PLoS One*, vol. 16, no. 3, e0247794, 2021.

[346] M. L. Reaka-Kudla, 'The global biodiversity of coral reefs: A comparison with rain forests,' *Biodiversity II: Understanding and protecting our biological resources*, vol. 2, p. 551, 1997.

[347] R. Revelle and H. E. Suess, 'Carbon dioxide exchange between atmosphere and ocean and the question of an increase of atmospheric co2 during the past decades,' *Tellus*, vol. 9, no. 1, pp. 18–27, 1957.

[348] B. Riegl, M. Johnston, S. Purkis, E. Howells, J. Burt, S. C. Steiner, C. R. Sheppard and A. Bauman, 'Population collapse dynamics in acropora downingi, an arabian/persian gulf ecosystem-engineering coral, linked to rising temperature,' *Global change biology*, vol. 24, no. 6, pp. 2447–2462, 2018.

[349] H. Ritchie and M. Roser, 'Forests and deforestation,' *Our World in Data*, 2021.

[350] C. M. Roberts, C. J. McClean, J. E. Veron, J. P. Hawkins, G. R. Allen, D. E. McAllister, C. G. Mittermeier, F. W. Schueler, M. Spalding, F. Wells *et al.*, 'Marine biodiversity hotspots and conservation priorities for tropical reefs,' *Science*, vol. 295, no. 5558, pp. 1280–1284, 2002.

[351] C. Roelfsema, E. Kovacs, P. Roos, D. Terzano, M. Lyons and S. Phinn, 'Use of a semi-automated object based analysis to map benthic composition, heron reef, southern great barrier reef,' *Remote Sensing Letters*, vol. 9, no. 4, pp. 324–333, 2018.

[352] C. M. Roelfsema, E. M. Kovacs, J. C. Ortiz, D. P. Callaghan, K. Hock, M. Mongin, K. Johansen, P. J. Mumby, M. Wettle, M. Ronan *et al.*, 'Habitat maps to enhance monitoring and management of the great barrier reef,' *Coral Reefs*, vol. 39, no. 4, pp. 1039–1054, 2020.

[353] C. M. Roelfsema, M. Lyons, N. Murray, E. M. Kovacs, E. Kennedy, K. Markey, R. Borrego-Acevedo, A. O. Alvarez, C. Say, P. Tudman *et al.*, 'Workflow for the generation of expert-derived training and validation data: A view to global scale habitat mapping,' *Frontiers in Marine Science*, 2021.

[354] J. S. Rogers, S. G. Monismith, R. B. Dunbar and D. Koweek, 'Field observations of wave-driven circulation over spur and groove formations on a coral reef,' *Journal of Geophysical Research: Oceans*, vol. 120, no. 1, pp. 145–160, 2015, ISSN: 21699291. DOI: 10.1002/2014JC010464.

[355] J. S. Rogers, S. G. Monismith, F. Feddersen and C. D. Storlazzi, 'Hydrodynamics of spur and groove formations on a coral reef,' *Journal of Geophysical Research: Oceans*, vol. 118, no. 6, pp. 3059–3073, 2013, ISSN: 21699291. DOI: 10.1002/jgrc.20225.

[356] F. Roth, F. Saalmann, T. Thomson, D. J. Coker, R. Villalobos, B. Jones, C. Wild and S. Carvalho, 'Coral reef degradation affects the potential for reef recovery after disturbance,' *Marine Environmental Research*, vol. 142, pp. 48–58, 2018.

[357] A. Rudiastuti, R. Dewi, Y. Ramadhani, A. Rahadiati, D. Sutrisno, W. Ambarwulan, I. Pujawati, E. Suryanegara, S. Wijaya, S. Hartini *et al.*, 'Benthic habitat mapping using sentinel 2a: A preliminary study in image classification approach in an absence of training data,' in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing, vol. 750, 2021, p. 012 029.

[358] C. L. Sabine, R. A. Feely, N. Gruber, R. M. Key, K. Lee, J. L. Bullister, R. Wanninkhof, C. Wong, D. W. Wallace, B. Tilbrook *et al.*, 'The oceanic sink for anthropogenic co2,' *science*, vol. 305, no. 5682, pp. 367–371, 2004.

[359] Y. Sadovy de Mitcheson, M. T. Craig, A. A. Bertoncini, K. E. Carpenter, W. W. Cheung, J. H. Choat, A. S. Cornish, S. T. Fennessy, B. P. Ferreira, P. C. Heemstra *et al.*, 'Fishing groupers towards extinction: A global assessment of threats and extinction risks in a billion dollar fishery,' *Fish and fisheries*, vol. 14, no. 2, pp. 119–136, 2013.

[360] K. Sambrook, A. S. Hoey, S. Andréfouët, G. S. Cumming, S. Duce and M. C. Bonin, 'Beyond the reef: The widespread use of non-reef habitats by coral reef fishes,' *Fish and Fisheries*, vol. 20, no. 5, pp. 903–920, 2019.

[361] A. H. Sanchez, M. C. A. Picoli, G. Camara, P. R. Andrade, M. E. D. Chaves, S. Lechler, A. R. Soares, R. F. Marujo, R. E. O. Simões, K. R. Ferreira *et al.*, 'Comparison of cloud cover detection algorithms on sentinel–2 images of the amazon tropical forest,' *Remote Sensing*, vol. 12, no. 8, p. 1284, 2020.

[362] A. Sandahl and A. P. Tøttrup, 'Marine citizen science: Recent developments and future recommendations,' *Citizen Science: Theory and Practice*, vol. 5, no. 1, 2020.

[363] E. Santos-Fernandez, E. E. Peterson, J. Vercelloni, E. Rushworth and K. Mengersen, 'Correcting misclassification errors in crowdsourced ecological data: A bayesian perspective,' *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 70, no. 1, pp. 147–173, 2021.

[364] P. P. Sarkar, P. Janardhan and P. Roy, 'Prediction of sea surface temperatures using deep learning neural networks,' *SN Applied Sciences*, vol. 2, no. 8, pp. 1–14, 2020.

[365] I. H. Sarker, 'Machine learning: Algorithms, real-world applications and research directions,' *SN computer science*, vol. 2, no. 3, p. 160, 2021.

[366] V. Sarukkai, A. Jain, B. Uzkent and S. Ermon, 'Cloud removal from satellite images using spatiotemporal generator networks,' in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1796–1805.

[367] J. Scealy and A. Welsh, 'Regression for compositional data by using distributions defined on the hypersphere,' *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 3, pp. 351–375, 2011.

[368] W. Schlager and S. Purkis, 'Reticulate reef patterns - antecedent karst versus self-organization,' *Sedimentology*, vol. 62, no. 2, pp. 501–515, 2015, ISSN: 13653091. DOI: 10.1111/sed.12172.

[369] J. Schmidhuber, 'Deep learning in neural networks: An overview,' *Neural networks*, vol. 61, pp. 85–117, 2015.

[370] V. Schoepf, S. A. Carrion, S. M. Pfeifer, M. Naugle, L. Dugal, J. Bruyn and M. T. McCulloch, 'Stress-resistant corals may not acclimatize to ocean warming but maintain heat tolerance under cooler temperatures,' *Nature communications*, vol. 10, no. 1, pp. 1–10, 2019.

[371] J. Scopélitis, S. Andréfouët, S. Phinn, P. Chabanet, O. Naim, C. Tourrand and T. Done, 'Changes of coral communities over 35 years: Integrating in situ and remote-sensing data on saint-leu reef (la réunion, indian ocean),' *Estuarine, Coastal and Shelf Science*, vol. 84, no. 3, pp. 342–352, 2009.

[372] J. Scopélitis, S. Andréfouët, S. Phinn, T. Done and P. Chabanet, 'Coral colonisation of a shallow reef flat in response to rising sea level: Quantification from 35 years of remote sensing data at heron island, australia,' *Coral Reefs*, vol. 30, no. 4, p. 951, 2011.

[373] E. Scornet, 'Tuning parameters in random forests,' *ESAIM: Proceedings and Surveys*, vol. 60, pp. 144–162, 2017.

[374] M. Segal-Rozenhaimer, A. Li, K. Das and V. Chirayath, 'Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (cnn),' *Remote Sensing of Environment*, vol. 237, p. 111 446, 2020.

[375] A. Sekulić, M. Kilibarda, G. B. Heuvelink, M. Nikolić and B. Bajat, 'Random forest spatial interpolation,' *Remote Sensing*, vol. 12, no. 10, p. 1687, 2020.

[376] M. B. Selamat, M. Lanuru and A. H. Muhiddin, 'Spatial composition of benthic substrate around bontosua island,' *Jurnal Ilmu Kelautan SPERMONDE*, vol. 4, no. 1, 2018.

[377] J. C. Selgrath, C. Roelfsema, S. E. Gergel and A. C. Vincent, 'Mapping for coral reef conservation: Comparing the value of participatory and remote sensing approaches,' *Ecosphere*, vol. 7, no. 5, e01325, 2016.

[378] A. Shapiro, D. Poursanidis, D. Traganos, L. Teixeira and L. Muaves, 'Mapping and monitoring the quirimbas national park seascape wwf-germany,' *Berlin, Germany*, 2020.

[379] Y. Shendryk, Y. Rist, C. Ticehurst and P. Thorburn, 'Deep learning for multi-modal classification of cloud, shadow and land cover scenes in planetscope and sentinel-2 imagery,' *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 157, pp. 124–136, 2019.

[380] P. Shi, A. Zhang and H. Li, 'Regression analysis for microbiome compositional data,' 2016.

[381] J. Shin, K. Kim and J.-H. Ryu, 'Comparative study on hyperspectral and satellite image for the estimation of chlorophyll a concentration on coastal areas,' *Korean Journal of Remote Sensing*, vol. 36, no. 2_2, pp. 309–323, 2020.

[382] Shinn, E., 'Spur and Groove Formation on the Florida Reef Tract,' *SEPM Journal of Sedimentary Research*, vol. Vol. 33, no. 2, pp. 291–303, 1963, ISSN: 1527-1404. DOI: 10.1306/74d70e34-2b21-11d7-8648000102c1865d.

[383] H. Shirmard, E. Farahbakhsh, R. D. Müller and R. Chandra, 'A review of machine learning in processing remote sensing data for mineral exploration,' *Remote Sensing of Environment*, vol. 268, p. 112 750, 2022.

[384] R. F. da Silva, C. D. Storlazzi, J. S. Rogers, J. Reyns and R. McCall, 'Modelling three-dimensional flow over spur-and-groove morphology,' *Coral Reefs*, vol. 39, no. 6, pp. 1841–1858, 2020.

[385] C. B. da Silveira, G. M. Strenzel, M. Maida, T. C. Araújo and B. P. Ferreira, 'Multiresolution satellite-derived bathymetry in shallow coral reefs: Improving linear algorithms with geographical analysis,' *Journal of Coastal Research*, vol. 36, no. 6, pp. 1247–1265, 2020.

[386] P. Singh and N. Komodakis, 'Cloud-gan: Cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks,' pp. 1772–1775, 2018.

[387] V. Siregar, S. Agus, A. Sunuddin, R. Pasaribu, M. Sangadji, A. Sugara and E. Kurniawati, 'Benthic habitat classification using high resolution satellite imagery in sebaru besar island, kepulauan seribu,' in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing, vol. 429, 2020, p. 012 040.

[388] W. Skirving, S. Enríquez, J. D. Hedley, S. Dove, C. M. Eakin, R. A. Mason, J. L. De La Cour, G. Liu, O. Hoegh-Guldberg, A. E. Strong *et al.*, 'Remote sensing of coral bleaching using temperature and light: Progress towards an operational algorithm,' *Remote Sensing*, vol. 10, no. 1, p. 18, 2018.

[389] A. L. Smith, D. M. Asta and C. A. Calder, 'The geometry of continuous latent space models for network data,' *Statistical science: a review journal of the Institute of Mathematical Statistics*, vol. 34, no. 3, p. 428, 2019.

[390] L. Smith, P. Cornillon, D. Rudnickas and C. B. Mouw, 'Evidence of environmental changes caused by chinese island-building,' *Scientific reports*, vol. 9, no. 1, pp. 1–11, 2019.

[391] Y. Song and H. Yan, 'Image segmentation techniques overview,' in *2017 Asia Modelling Symposium (AMS)*, IEEE, 2017, pp. 103–107.

[392] C. Sothe, C. De Almeida, M. Schimalski, V. Liesenberg, L. La Rosa, J. Castro and R. Feitosa, 'A comparison of machine and deep-learning algorithms applied to multisource data for a subtropical forest area classification,' *International Journal of Remote Sensing*, vol. 41, no. 5, pp. 1943–1969, 2020.

[393] D. Sous, G. Dodet, F. Bouchette and M. Tissier, 'Momentum Balance Across a Barrier Reef,' *Journal of Geophysical Research: Oceans*, vol. 125, no. 2, pp. 1–24, 2020, ISSN: 21699291. DOI: 10.1029/2019JC015503.

[394] D. Sous, F. Bouchette, E. Doerflinger, S. Meulé, R. Certain, G. Toulemonde, B. Dubarbier and B. Salvat, 'On the small-scale fractal geometrical structure of a living coral reef barrier,' *Earth Surface Processes and Landforms*, vol. 45, no. 12, pp. 3042–3054, 2020.

[395] D. Sous, C. Chevalier, J. L. Devenon, J. Blanchot and M. Pagano, 'Circulation patterns in a channel reef-lagoon system, Ouano lagoon, New Caledonia,' *Estuarine, Coastal and Shelf Science*, vol. 196, pp. 315–330, 2017, ISSN: 02727714. DOI: 10.1016/j.ecss.2017.07.015.

[396] D. Sous, G. Dodet, F. Bouchette and M. Tissier, 'Momentum balance across a barrier reef,' *Journal of Geophysical Research: Oceans*, vol. 125, no. 2, e2019JC015503, 2020.

[397] D. Sous, K. Martins, M. Tissier, F. Bouchette and S. Meulé, 'Spectral wave dissipation over a roughness-varying barrier reef,' *Geophysical Research Letters*, vol. 50, no. 5, e2022GL102104, 2023.

[398] D. Sous, S. Maticka, S. Meulé and F. Bouchette, 'Bottom drag coefficient on a shallow barrier reef,' *Geophysical Research Letters*, vol. 49, no. 6, e2021GL097628, 2022.

[399] D. Sous, M. Tissier, V. Rey, J. Touboul, F. Bouchette, J.-L. Devenon, C. Chevalier and J. Aucan, 'Wave transformation over a barrier reef,' *Continental Shelf Research*, vol. 184, pp. 66–80, 2019.

[400] M. Spalding, C. Ravilious and E. P. Green, *World atlas of coral reefs*. Univ of California Press, 2001.

[401] M. Spalding and A. Grenfell, 'New estimates of global and regional coral reef areas,' *Coral reefs*, vol. 16, no. 4, pp. 225–230, 1997.

[402] I. R. Staude, E. Vélez-Martin, B. O. Andrade, L. R. Podgaiski, I. I. Boldrini, M. Mendonca Jr, V. D. Pillar and G. E. Overbeck, 'Local biodiversity erosion in south brazilian grasslands under moderate levels of landscape habitat loss,' *Journal of Applied Ecology*, vol. 55, no. 3, pp. 1241–1251, 2018.

[403] R. D. Stuart-Smith, G. J. Edgar, N. S. Barrett, A. E. Bates, S. C. Baker, N. J. Bax, M. A. Becerro, J. Berkhout, J. L. Blanchard, D. J. Brock *et al.*, 'Assessing national biodiversity trends for rocky and coral reefs through the integration of citizen science and scientific monitoring programs,' *Bioscience*, vol. 67, no. 2, pp. 134–146, 2017.

[404] J. Stuckens, P. Coppin and M. Bauer, 'Integrating contextual information with per-pixel classification for improved land cover classification,' *Remote sensing of environment*, vol. 71, no. 3, pp. 282–296, 2000.

[405] S. Sully, D. Burkepile, M. Donovan, G. Hodgson and R. Van Woesik, 'A global analysis of coral bleaching over the past two decades,' *Nature communications*, vol. 10, no. 1, pp. 1–5, 2019.

[406] D. Sutrisno, A. Sugara and M. Darmawan, 'The assessment of coral reefs mapping methodology: An integrated method approach,' in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing, vol. 750, 2021, p. 012 030.

[407] S. Talukdar, P. Singha, S. Mahato, S. Pal, Y.-A. Liou and A. Rahman, 'Land-use land-cover classification by machine learning classifiers for satellite observations—a review,' *Remote Sensing*, vol. 12, no. 7, p. 1135, 2020.

[408] N. Tatar, M. Saadatseresht, H. Arefi and A. Hadavand, 'A robust object-based shadow detection method for cloud-free high resolution satellite images over urban areas and water bodies,' *Advances in Space Research*, vol. 61, no. 11, pp. 2787–2800, 2018.

[409] *Tecator meat sample dataset*, http://lib.stat.cmu.edu/datasets/tecator.

[410] A. P. Tewkesbury, A. J. Comber, N. J. Tate, A. Lamb and P. F. Fisher, 'A critical synthesis of remotely sensed optical image change detection techniques,' *Remote Sensing of Environment*, vol. 160, pp. 1–14, 2015.

[411] P. Thanh Noi and M. Kappas, 'Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery,' *Sensors*, vol. 18, no. 1, p. 18, 2018.

[412] G. W. Thum, S. H. Tang, S. A. Ahmad and M. Alrifaey, 'Toward a highly accurate classification of underwater cable images via deep convolutional neural network,' *Journal of Marine Science and Engineering*, vol. 8, no. 11, p. 924, 2020.

[413] Z. Tian, J. Zhu and B. Han, 'Research on coral reefs monitoring using worldview-2 image in the xiasha islands,' in *Second Target Recognition and Artificial Intelligence Summit Forum*, International Society for Optics and Photonics, vol. 11427, 2020, 114273S.

[414] L. Torgo, P. Branco, R. P. Ribeiro and B. Pfahringer, 'Resampling strategies for regression,' *Expert Systems*, vol. 32, no. 3, pp. 465–476, 2015.

[415] L. Torgo, R. P. Ribeiro, B. Pfahringer and P. Branco, 'Smote for regression,' in *Portuguese conference on artificial intelligence*, Springer, 2013, pp. 378–389.

[416] D. Traganos and P. Reinartz, 'Machine learning-based retrieval of benthic reflectance and posidonia oceanica seagrass extent using a semi-analytical inversion of sentinel-2 satellite data,' *International Journal of Remote Sensing*, vol. 39, no. 24, pp. 9428–9452, 2018.

[417] D.-C. Tseng, H.-T. Tseng and C.-L. Chien, 'Automatic cloud removal from multi-temporal spot images,' *Applied Mathematics and Computation*, vol. 205, no. 2, pp. 584–600, 2008.

[418] M. C. Tsilimigras and A. A. Fodor, 'Compositional data analysis of the microbiome: Fundamentals, tools, and challenges,' *Annals of epidemiology*, vol. 26, no. 5, pp. 330–335, 2016.

[419] Y.-H. Tu, S. Phinn, K. Johansen and A. Robson, 'Assessing radiometric correction approaches for multi-spectral uas imagery for horticultural applications,' *Remote Sensing*, vol. 10, no. 11, p. 1684, 2018.

[420] H. Tuomisto *et al.*, 'Influence of compositing criterion and data availability on pixel-based landsat tm/etm+ image compositing over amazonian forests,' *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 3, pp. 857–867, 2016.

[421] UNESCO World Heritage Convention, *Great barrier reef*, [Online; accessed 25-August-2023]. [Online]. Available: https://whc.unesco.org/en/list/154/.

[422] M. C. Urban, G. Bocedi, A. P. Hendry, J.-B. Mihoub, G. Pe'er, A. Singer, J. Bridle, L. Crozier, L. De Meester, W. Godsoe *et al.*, 'Improving the forecast for biodiversity under climate change,' *Science*, vol. 353, no. 6304, aad8466, 2016.

[423] J. Van Den Bergh, V. Chirayath, A. Li, J. L. Torres-Pérez and M. Segal-Rozenhaimer, 'Nemo-net–gamifying 3d labeling of multi-modal reference datasets to support automated marine habitat mapping,' *Frontiers in Marine Science*, vol. 8, p. 347, 2021.

[424] A. Van Dongeren, R. Lowe, A. Pomeroy, D. M. Trang, D. Roelvink, G. Symonds and R. Ranasinghe, 'Numerical modeling of low-frequency wave dynamics over a fringing coral reef,' *Coastal Engineering*, vol. 73, pp. 178–190, 2013, ISSN: 03783839. DOI: 10.1016/j.coastaleng.2012.11.004.

[425] R. Van Klink, D. E. Bowler, K. B. Gongalsky, A. B. Swengel, A. Gentile and J. M. Chase, 'Meta-analysis reveals declines in terrestrial but increases in freshwater insect abundances,' *Science*, vol. 368, no. 6489, pp. 417–420, 2020.

[426] S. Van Wynsberge, C. Menkes, R. Le Gendre, T. Passfield and S. Andréfouët, 'Are sea surface temperature satellite measurements reliable proxies of lagoon temperature in the south pacific?' *Estuarine, Coastal and Shelf Science*, vol. 199, pp. 117–124, 2017.

[427] A. Vedaldi and S. Soatto, 'Quick shift and kernel methods for mode seeking,' in *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV 10*, Springer, 2008, pp. 705–718.

[428] J. Vercelloni, M. Kayal, Y. Chancerelle and S. Planes, 'Exposure, vulnerability, and resiliency of french polynesian coral reefs to environmental disturbances,' *Scientific Reports*, vol. 9, no. 1, pp. 1–11, 2019.

[429] J. Vercelloni, B. Liquet, E. V. Kennedy, M. González-Rivero, M. J. Caley, E. E. Peterson, M. Puotinen, O. Hoegh-Guldberg and K. Mengersen, 'Forecasting intensifying disturbance effects on coral reefs,' *Global Change Biology*, vol. 26, no. 5, pp. 2785–2797, 2020.

[430] M. B. Villanueva and M. A. Ballera, 'Multinomial classification of coral species using enhanced supervised learning algorithm,' in *2020 IEEE 10th International Conference on System Engineering and Technology (ICSET)*, IEEE, 2020, pp. 202–206.

[431] N. Wahidin, V. P. Siregar, B. Nababan, I. Jaya and S. Wouthuyzen, 'Object-based image analysis for coral reef benthic habitat mapping with several classification algorithms,' *Procedia Environmental Sciences*, vol. 24, no. Supplement C, pp. 222–227, 2015.

[432] J. Wan and Y. Ma, 'Multi-scale spectral-spatial remote sensing classification of coral reef habitats using cnn-svm,' *Journal of Coastal Research*, vol. 102, no. SI, pp. 11–20, 2020.

[433] X. Wang, Y. Zhong, L. Zhang and Y. Xu, 'Blind hyperspectral unmixing considering the adjacency effect,' *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6633–6649, 2019.

[434] Y. Wei, Z. Wang, H. Wang, Y. Li and Z. Jiang, 'Predicting population age structures of china, india, and vietnam by 2030 based on compositional data,' *PLoS One*, vol. 14, no. 4, e0212772, 2019.

[435] Y. Wei, Z. Wang, H. Wang, T. Yao and Y. Li, 'Promoting inclusive water governance and forecasting the structure of water consumption based on compositional data: A case study of beijing,' *Science of the Total Environment*, vol. 634, pp. 407–416, 2018.

[436] P. Wicaksono, 'Improving the accuracy of multispectral-based benthic habitats mapping using image rotations: The application of principle component analysis and independent component analysis,' *European Journal of Remote Sensing*, vol. 49, no. 1, pp. 433–463, 2016.

[437] P. Wicaksono and P. A. Aryaguna, 'Analyses of inter-class spectral separability and classification accuracy of benthic habitat mapping using multispectral image,' *Remote Sensing Applications: Society and Environment*, vol. 19, p. 100 335, 2020.

[438] P. Wicaksono, P. A. Aryaguna and W. Lazuardi, 'Benthic habitat mapping model and cross validation using machine-learning classification algorithms,' *Remote Sensing*, vol. 11, no. 11, p. 1279, 2019.

[439] P. Wicaksono and W. Lazuardi, 'Assessment of planetscope images for benthic habitat and seagrass species mapping in a complex optically shallow water environment,' *International journal of remote sensing*, vol. 39, no. 17, pp. 5739–5765, 2018.

[440] P. Wicaksono and W. Lazuardi, 'Random forest classification scenarios for benthic habitat mapping using planetscope image,' pp. 8245–8248, 2019.

[441] G. Wilkinson and J. Megier, 'Evidential reasoning in a pixel classification hierarchy—a potential method for integrating image classifiers and expert system rules based on geographic context,' *Remote Sensing*, vol. 11, no. 10, pp. 1963–1968, 1990.

[442] M. J. Williamson, E. J. Tebbs, T. P. Dawson, H. J. Thompson, C. E. Head and D. M. Jacoby, 'Monitoring shallow coral reef exposure to environmental stressors using satellite earth observation: The reef environmental stress exposure toolbox (reset),' *Remote Sensing in Ecology and Conservation*, vol. 8, no. 6, pp. 855–874, 2022.

[443] D. L. Wilson, 'Asymptotic properties of nearest neighbor rules using edited data,' *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-2, no. 3, pp. 408–421, 1972.

[444] K. Wolfe, K. Anthony, R. C. Babcock, L. Bay, D. G. Bourne, D. Burrows, M. Byrne, D. J. Deaker, G. Diaz-Pulido, P. R. Frade *et al.*, 'Priority species to support the functional integrity of coral reefs,' *Oceanography and Marine Biology*, 2020.

[445] S. Wouthuyzen, M. Abrar, C. Corvianawatie, A. Salatalohi, S. Kusumo, Y. Yanuar, M. Arrafat *et al.*, 'The potency of sentinel-2 satellite for monitoring during and after coral bleaching events of 2016 in the some islands of marine recreation park (twp) of pieh, west sumatra,' vol. 284, no. 1, p. 012 028, 2019.

[446] F. Xia, J. Chen, W. K. Fung and H. Li, 'A logistic normal multinomial regression model for microbiome compositional data analysis,' *Biometrics*, vol. 69, no. 4, pp. 1053–1063, 2013.

[447] H. Xu, Z. Liu, J. Zhu, X. Lu and Q. Liu, 'Classification of coral reef benthos around ganquan island using worldview-2 satellite imagery,' *Journal of Coastal Research*, vol. 93, no. SI, pp. 466–474, 2019.

[448] Y. Xu, N. R. Vaughn, D. E. Knapp, R. E. Martin, C. Balzotti, J. Li, S. A. Foo and G. P. Asner, 'Coral bleaching detection in the hawaiian islands using spatio-temporal standardized bottom reflectance and planet dove satellites,' *Remote Sensing*, vol. 12, no. 19, p. 3219, 2020.

[449] H. F. Yan, P. M. Kyne, R. W. Jabado, R. H. Leeney, L. N. Davidson, D. H. Derrick, B. Finucci, R. P. Freckleton, S. V. Fordham and N. K. Dulvy, 'Overfishing and habitat loss drive range contraction of iconic marine fishes to near extinction,' *Science Advances*, vol. 7, no. 7, eabb6026, 2021.

[450] W. Y. Yan and A. Shaker, 'Radiometric correction and normalization of airborne lidar intensity data for improving land-cover classification,' *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 12, pp. 7658–7673, 2014.

[451] R. Yanovski and A. Abelson, 'Structural complexity enhancement as a potential coral-reef restoration tool,' *Ecological Engineering*, vol. 132, pp. 87–93, 2019.

[452] M. Yasir, A. U. Rahman and M. Gohar, 'Habitat mapping using deep neural networks,' *Multimedia Systems*, pp. 1–12, 2020.

[453] J.-M. Yeom, J.-L. Roujean, K.-S. Han, K.-S. Lee and H.-W. Kim, 'Thin cloud detection over land using background surface reflectance based on the brdf model applied to geostationary ocean color imager (goci) satellite data sets,' *Remote Sensing of Environment*, vol. 239, p. 111 610, 2020.

[454] L. Yu and A. Ettinger, 'Assessing phrasal representation and composition in transformers,' *arXiv preprint arXiv:2010.03763*, 2020.

[455] J. Zalasiewicz*, M. Williams, W. Steffen and P. Crutzen, *The new world of the anthropocene*, 2010.

[456] P. A. Zapata-Ramírez, P. Blanchon, A. Olioso, H. Hernandez-Nuñez and J. A. Sobrino, 'Accuracy of ikonos for mapping benthic coral-reef habitats: A case study from the puerto morelos reef national park, mexico,' *International Journal of Remote Sensing*, vol. 34, no. 9-10, pp. 3671–3687, 2013.

[457] Z. Zhafarina and P. Wicaksono, 'Benthic habitat mapping on different coral reef types using random forest and support vector machine algorithm,' in *Sixth International Symposium on LAPAN-IPB Satellite*, International Society for Optics and Photonics, vol. 11372, 2019, p. 113721M.

[458] X. Zhang, P. Xiao and X. Feng, 'Impervious surface extraction from high-resolution satellite image using pixel-and object-based hybrid analysis,' *International journal of remote sensing*, vol. 34, no. 12, pp. 4449–4465, 2013.

[459] Z. Zhang, G. He and X. Wang, 'A practical dos model-based atmospheric correction algorithm,' *International Journal of Remote Sensing*, vol. 31, no. 11, pp. 2837–2852, 2010.

[460] Z. Zhou, L. Ma, T. Fu, G. Zhang, M. Yao and M. Li, 'Change detection in coral reef environment using high-resolution images: Comparison of object-based and pixel-based paradigms,' *ISPRS International Journal of Geo-Information*, vol. 7, no. 11, p. 441, 2018.

[461] X. Zhu and E. H. Helmer, 'An automatic method for screening clouds and cloud shadows in optical satellite image time series in cloudy regions,' *Remote sensing of environment*, vol. 214, pp. 135–153, 2018.

[462] Z. Zhu, 'Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications,' *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 130, pp. 370–384, 2017.

[463] Z. Zhu and C. E. Woodcock, 'Object-based cloud and cloud shadow detection in landsat imagery,' *Remote sensing of environment*, vol. 118, pp. 83–94, 2012.

[464] M. L. Zoffoli, R. Frouin and M. Kampel, 'Water column correction for coral reef studies by remote sensing,' *Sensors*, vol. 14, no. 9, pp. 16 881–16 931, 2014.