Automatic geomorphological mapping using ground truth data with coverage sampling and random forest algorithms

Paul Aimé Latsouck Faye, Elodie Brunel, Thomas Claverie, Solym Mawaki Manou-Abi, Sophie Dabo-Niang

Automatic geomorphological mapping using ground truth data with coverage sampling and random forest algorithms

Faye Paul Aimé Latsouck[1*], Brunel Elodie[1†], Claverie Thomas[2†], Manou-Abi Solym[1,3,5†], Dabo-Niang Sophie[4†]

[1] IMAG, University of Montpellier, CNRS, 34090, Montpellier, France .
[2] UMR ENTROPIE, IRD, IFREMER, CNRS, Univ La Réunion, Saint Denis, 97744, Réunion, France .
[3] CUFR, Centre Universitaire de Formation et de Recherche, Dembéni, 97660, Mayotte, France .
[4] PAINLEVE UMR 8524, University of Lille, CNRS, INRIA-MODAL, 59665, Lille, France .
[5] Laboratoire de Mathématiques et Applications UMR 7348 , University of Poitiers, CNRS, Futuroscope, 86073, Poitiers, France .

*Corresponding author(s). E-mail(s): paul-aime-latsouck.faye@umontpellier.fr;
Contributing authors: elodie.brunel-piccinini@umontpellier.fr; tclaverie@gmail.com; solym.manou-abi@univ-mayotte.fr; sophie.dabo@univ-lille.fr;
†These authors contributed equally to this work.

Abstract

Marine geomorphological maps are useful to understand seafloor structure for example in the context of ecological studies, resources management or conservation planning. Although techniques to build such maps are increasingly sophisticated, manual techniques are still largely used. Automated approaches are needed to get reproducible maps in a reasonable time. This work provides statistical learning approaches based framework to build automatically geomorphological maps. We used bathymetric data to build Digital Bathymetric Model (DBM) and compute terrain attributes characteristic of seafloor geomorphology. Then, we used clustering based algorithms to select automatically ground

truth locations from a reference geomorphological map manually made. Finally a supervised classification model random forest based was used to build predictive models for seafloor geomorphology typologies. Subsequently we studied the effect of DBM resolution, sample size and sampling method of the ground truth locations, in the quality of map production via a series of simulations. Results showed that the proposed framework allowed to build efficiently relevant seafloor geomorphological maps. The best compromise between the sampling effort and the quality of the resulting maps was obtained with 100 m DBM resolution, 200 data points sample size and using a complexity-dependent sampling method and led to a map matching at 90% the reference one.

Keywords: geomorphological map, spatial modeling, random forest classification, digital bathymetric model, terrain attributes, lidar data

Declarations

Compliance with Ethical Standards

All accepted principles of ethical and professional conduct have been followed during this research in accordance with Springer's standards.

Funding

No funding was received for conducting this study.

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical Approval

This manuscript has not been published, accepted for publication, or under editorial review for publication elsewhere.

Informed Consent

The authors have given the informed consent to publish this article in the Journal of Geovisualization and Spatial Analysis if accepted.

Code Availibility

The R code created during this work is open source and can be accessed in Zenodo repositories : https://doi.org/10.5281/zenodo.8436795

Aknowledgments

EPICURE program provides the bathymetric data used in this project. We are thankful to Rodolphe Devillers and Christophe Crambes for their field support. We thank Baptiste Chapuisat, Ghislain Durif and François David Collain for their assistance in High Performance Computing techniques. We also thank meso@LR for computational resources provided for the simulations carried out in this work.

1 Introduction

Geomorphological maps are georeferenced delineation of morphological structure and surface composition of a studied land and/or seafloor (Otto and Smith, 2013; Dramis et al, 2011; Pavlopoulos et al, 2009). Marine geomorphological maps are particularly crucial for resource management, conservation efforts, hazard assessment, protected area management, effective marine research campaign planning and various marine-related industrial works (Kienholz, 1978; Bishop et al, 2012; Fukunaga et al, 2019; Browne et al, 2010). Such maps provide valuable information on the composition and structure of the seabed which, among other usage, is needed to generate habitat maps through the identification and delineation of different benthic ecosystems like coral reefs, seagrass beds or various deep-sea communities (Pandian et al, 2009; Dramis et al, 2011; Wabnitz et al, 2008).

To produce geomorphological maps, different approaches are used (Siart et al, 2009; Hugenholtz et al, 2013). Widely used imagery techniques involve studying the patterns, textures, shapes, and color variations present in the imagery. This can be done manually by digitizing or tracing the features on the imagery or through automated or semi-automated image segmentation and classification techniques in Geographical Information System (GIS) using available image analysis tools. While imagery can be a valuable tool, it has a limited use for mapping deeply submerged geomorphological features due to water opacity. In such conditions, only acoustic approaches can provide usable data which are generally completed by punctual carefully located ground truthing observations using for example scuba-divers or submarine-divers observation, Remote Operated Vehicle (ROV) or Automatic Underwater Vehicle (AUV) picture or videos, drop cameras or seabed sampling (Wynn et al, 2014; Locker et al, 2010). Subsequent treatments, required to generate maps with such data, will be to first propose a geomorphologic category for each ground truthing point (ie. Typology), then delineate surfaces of homogeneous typologies. Depending on whether ground truthing points are defined as categories or as quantitative data, interpolation methodologies within the surface to be mapped might take different forms.

Many semi-automated or automated approaches have been also proposed in recent decades to achieve objective, automated and repeatable approach to extract meaningful information using vast quantities of data (Summers et al, 2021). Object-Based Image Analysis (OBIA) which use bathymetric derivatives or a combination of bathymetric derivatives and backscatter to automatically segment the seafloor (Masetti et al, 2018; Argyropoulou et al, 2016; Lacharité et al, 2018; Koop et al, 2021; Dekavalla and Argialas, 2017) are widely used. Bathymorphon / geomorphon-based classification (Jasiewicz and Stepinski, 2013; Sowers et al, 2020; Ahn et al, 2023; Novaczek et al, 2019) and fuzzy logic scheme (Schmidt and Hewitt, 2004; Lucieer and Lucieer, 2009; Janowski et al, 2021) have been also investigated. The past 10 years, machine learning (Maschmeyer et al, 2019; Misiuk et al, 2021; Janowski et al, 2022; Sklar et al, 2024) and deep learning models (Behrens et al, 2018; Azarafza et al, 2023; Arhant et al, 2023) has increasingly been used for geomorphological mapping. Furtherthemore, the performance of these statistical learning methods has also been investigated in comparison with manually ones (Van der Meij et al, 2022; Diesing et al, 2014). In recent years, although these methods have proven their effectiveness, they are often combined

with underwater imagery which superseded expert manual interpretation and are particularly costly for large scale mapping (Van der Meij et al, 2022; Cui et al, 2021; Galvez et al, 2022; Misiuk and Brown, 2023; Breyer et al, 2023). As alternative, this work provides a clustering based algorithm for an optimal ground truth sampling and a learning-based approach for automatic geomorphological mapping. It focuses on the classical situation where experts use morphological map to define surfaces and ground truth measure to define typologies. But to that respect, further considerations need to be addressed: (i) the quality of bathymetric data required, (ii) the typology definition, and (iii) the methodology to use for creating geomorphological maps with such data.
(i) Bathymetry is recognized as essential when mapping geomorphology (Wilson et al, 2007; Lecours et al, 2016; Fukunaga et al, 2019). It is particularly useful for identifying and delineating submerged landforms that are not visible in aerial images. Modern bathymetric systems can provide high-resolution data, capturing fine-scale details of the seafloor morphology. Their accuracy is often higher compared to aerial images, as it directly measures the water depth and seafloor elevation. However, the selection of the appropriate technology is crucial to ensure the acquisition of accurate and reliable data suited for an optimal geomorphological map production. More precisely equipment and material setup will depend on factors such as the scale and coverage needed, water depth, desired resolution, sampling effort and budgetary considerations. Compromise between the sampling effort and the quality of the data have to be made but little tools are available to choose the most optimal setup. Indeed very high-resolution data may be necessary for detailed local studies, while coarser resolution data might suffice for larger-scale regional or global analyses. In the present research, some examples to help on such decisions are proposed.
(ii) A typology is a description of one geomorphological category of seabed (in our case). However, to make a geomorphological map usable, comparable and understandable by a large international community, categories definition of typologies need to form a consensus. Our study is based on the Millennium Coral Reef Mapping Project (MCRMP) typologies. Initiated by the Institute for Marine Remote Sensing - University of South Florida (IMaRS/USF) in 2001 and continued since 2003 by reasearchers of the Institut de Recherche pour le Développement (IRD), the MCRMP has proposed a multi-level hierarchical structure (Andréfouët et al, 2004). MCRMP typologies were widely used as they enable several sites around the world to be compared on a thematically rich, homogeneous basis (Andréfouët and Dirberg, 2006). These typologies provide a description of coral reef geomorphology distinguishing reef units such as slopes, flats, passes, terraces, lagoons, channels, etc.
(iii) Several approaches can be chosen to generate geomorphological maps from previously described data. The most classical approach is to manually draw typologies envelopes using GIS softwares (Minár and Evans, 2008; Otto et al, 2018). However, it is tedious and poorly replicable if temporal changes need to be monitored. Automated approaches are of several kind. This work focuses on statistical learning based approaches for their ability to efficiently process and analyze large volumes of spatial data, to learn complex relationships present in the data leading to improved accuracy in produced maps (Stepinski et al, 2007; Siqueira et al, 2022; Van der Meij et al, 2022). More precisely four steps are followed : 1) generate a Digital Bathymetric Model

(DBM) from bathymetric data, 2) compute terrain attributes from DBM on the entire surface studied, 3) train random forest based classification algorithm to match terrain attributes with ground truthing typologies and 4) predict typologies from the entire surface studied using the classification model generated.
The chosen methodology to generate geomorphological map was tested on a classical tropical reef feature: an atoll mapping from South Western Indian Ocean. Our approach is to propose a sensitivity analysis based on a comparison with an existing expert map while degrading DBM definition as proxy of data acquisition setup and varying the number of ground truth points as well as the methodology to select their locations as proxy of field effort. The ultimate goal is to supply tools to plan mapping field campaign using coverage sampling algorithms, codes to semi-automate geomorphological mapping procedure using random forest algorithm and propose metrics to evaluate the quality of the map generated across different resolutions. Recommendations to choose the DBM resolution, ground truth size and sites selection are also provided.

2 Materials and methods

2.1 Data

For this work, bathymetric data (depth measurements) collected between 2009 and 2010 on Geyser atoll with a surface area of approximately 268 km^2 are used (Figure 1A). A set of 48.10^6 data points were collected by LIDAR 1 m resolution calibrated. The depth range captured by this tool often reaches down to -30 m but due to exceptional very good water clarity conditions, bathymetric records on Geyser range between -50 and 4 meters (see their distribution in Figure S1). This data is available on the Hydrography and Oceanography Service of the Navy website [1].
An expert geomorphological map of the Geyser atoll available here [2] is also used. This scale-free map was produced by manual contouring, resulting from expert interpretation of hyperspectral images. 11 geomorphological typologies have been identified on Geyser (Roos et al, 2017).

2.2 General methodology

Here an automated scheme using bathymetry and some field typologies verifications also called ground truth to generate reproducible geomorphological maps is used (see Figure S2 in the supplemental materials). The methodology is the following: 1) Using bathymetric data, a Digital Bathymetric Model (DBM) is created at a given resolution. 2) From the DBM, terrain attributes corresponding to raster layers which have the potential to influence seafloor geomorphology are computed. 3) Using a given coordinate sampling method, a given number of ground truth locations is drawn. By overlaying these locations with the expert map, geomorphological typologies are attributed to each ground truth points. 4) Using the training set of coupled terrain attributes and ground truth typologies, a recursive feature elimination algorithm based

[1] https://data.shom.fr/donnees
[2] https://sextant.ifremer.fr/geonetwork/srv/api/records/21232c12-e409-4136-a24a-78c346518cfa

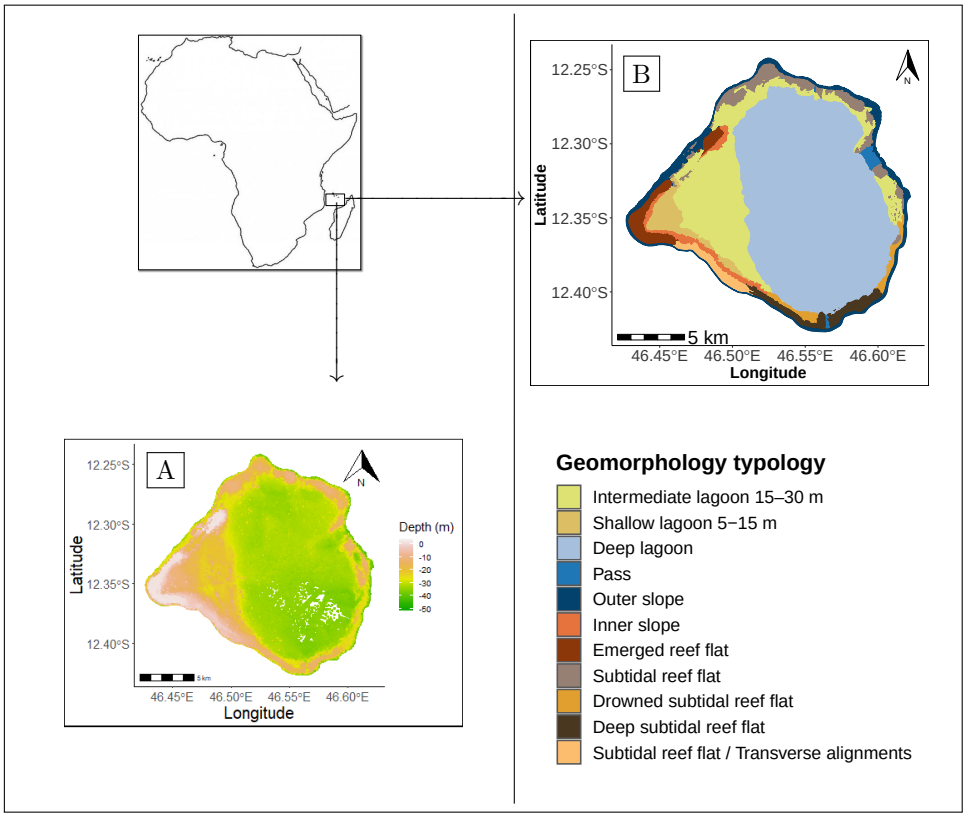


Fig. 1 Geyser is an atoll in the Western Indian ocean, specifically located in the northern Mozambique channel between Mayotte and the Gloriosos Islands and covering approximately 268 km^2. **A** Bathymetry data (in meters) collected by LIDAR technology and provided by Hydrography and Oceanography Service of the Navy (SHOM). **B** Geomorphologic structure identified for the Geyser atoll, as a part of the EPICURE project (Roos et al, 2017).

on a random forest classifier is used to select most relevant covariates. The latter are then used to train a random forest based predictive model for geomorphologic typologies. This model is finally used to predict geomorphologic typologies on the whole study site. 5) The quality of the map production was then evaluated using two types of performance criteria: the model performance (Balanced accuracy) and the matching between the generated map and the expert map (Match and Balanced match).
The robustness of our approach is also evaluated using a sentivity analysis through simulation by varying bathymetric data definition, numbers of ground truth points and methodology to select their locations. This involved in examining five DBM resolutions (5 m, 25 m, 50 m, 100 m and 500 m), six sample sizes (50, 100, 200, 500, 700, 1000) and three sampling methods (two spatial coverage sampling methods denoted SCS-KMEANS and SCS-CLARA, and a complexity-dependent sampling method denoted CDS). For each of these combinations ($5 \times 6 \times 3$), 30 replicates of geomorphological maps are generated by replicating the steps 3-5 described above.

2.2.1 Digital Bathymetric Model creation

The raw bathymetric dataset contains approximately 48.10^6 data points giving latitude, longitude and depth values. To avoid numerical issues due to redundant observations, this dataset was spatially sub-sampled and data points are kept as homogeneous as possible using the $buffer.points$ function in the supplemental materials (Roberts, 2015). The sub-sampled data are such that each retained data point is at least at a distance of 5 m from one another. Following this procedure, the new dataset of approximately 8.10^6 data points obtained is used to generate the DBM which consists in the creation of a square grid of T^2 cells or rasters. The cell size define the DBM resolution. Different resolutions (5 m, 25 m, 50 m, 100 m, 500 m)[3] are created using the *dbm* function proposed in the supplemental materials. The bathymetry or depth value of each cell is calculated by taking the average of all the depths inside the same cell.

Furthermore, some deep areas are poorly sampled, due to the lack of signal return on the LIDAR sensor (see white zone in the middle of Figure 1A). To get a depth measure for each ground truth location selected using one of the automatic sampling methodology presented further, depth measure on each raster of the study area were required. Thus for each DBM resolution, missing depth of empty cells in this zone are interpolated using an ordinary kriging model via the *ok.dbm* function provided in the supplemental materials. The kriging method is not decribed in details since the problem of missing data is out of the scope of the paper (Cressie, 1988). In contrast, empirical results provided in the supplemental materials show that it performs better than five others spatial interpolation methods for our data Table S1. This imputation method allows to complete bathymetric data on these cells.

2.2.2 Terrain attributes calculation

It consists in quantifying predictors for seafloor geomorphology. For each DBM resolution, terrain attributes were calculated using a moving routine from the DBM. More precisely, let us denote the depth $D_{i,j}$ of a given cell with $(i,j) \in [\![1,T]\!] \times [\![1,T]\!]$ and consider the depth $D_{i+k,j+l}$ of the neighboring cells with $(k,l) \in \mathcal{D}_3 = [\![-1,1]\!] \times [\![-1,1]\!]$. These cells are defining the 3×3 window on which the terrain attributes are defined. Note that it is possible to generalize the study to square window with size greater than 3. Selected terrain attributes can be organised into three groups: Slope ($Slope$) and orientation ($Aspect$) measures, terrain variability measures such as Roughness ($Roughness$), Terrain ruggedness index (TRI) and Vector ruggedness measure (VRM) and curvature and relative position measures such as Profile convexity ($profc$), Planform convexity ($planc$) and Bathymetric position index (BPI).

Slope has been widely recognized as an important factor for determining benthic habitat and colonization and has been used in many marine studies (Copeland et al, 2013; Fukunaga et al, 2019; Sterne et al, 2020). Like the magnitude of the steepest drop

[3] The 5 m resolution is the one used by biologists during field verification campaigns. We then looked for a very high resolution at which the geomorphological maps produced were degraded because they were too pixelated. 500 m seemed to be a good choice. Between these two resolutions, we empirically searched for intermediate resolutions enabling us to obtain "significantly" different maps. Hence the 25 m, 50 m and 100 m resolutions were retained.

in depth, *Slope* (in degrees) is derived from rates of change in x (longitude) and y (latitude) directions and is calculated as follows, for each cell $(i,j) \in [\![1,T]\!] \times [\![1,T]\!]$,

$$Slope_{i,j} = \frac{180}{\pi} \arctan\left(\sqrt{\left(\frac{\partial D_{i,j}}{\partial x}\right)^2 + \left(\frac{\partial D_{i,j}}{\partial y}\right)^2}\right). \tag{1}$$

where

$$\begin{cases} \dfrac{\partial D_{i,j}}{\partial x} = \big[(D_{i+1,j+1} + 2D_{i+1,j} + D_{i+1,j-1}) - (D_{i-1,j+1} + 2D_{i-1,j} + D_{i-1,j-1})\big]/8\Delta. \\ \dfrac{\partial D_{i,j}}{\partial y} = \big[(D_{i+1,j+1} + 2D_{i,j+1} + D_{i-1,j+1}) - (D_{i+1,j-1} + 2D_{i,j-1} + D_{i-1,j-1})\big]/8\Delta. \end{cases}$$

where the real number Δ stands for the cell size of the grid.
The orientation measure (*Aspect*) gives the exposure of a given area to such water waves and is often used in the calculation of others parameters that directly influence habitat (Wilson et al, 2007). *Aspect* (in degrees) is the compass direction of the steepest drop in depth and is calculated as follows, for each cell $(i,j) \in [\![1,T]\!] \times [\![1,T]\!]$,

$$Aspect_{i,j} = 180 + \frac{180}{\pi} \arctan\left(\frac{\partial D_{i,j}}{\partial x} + \frac{\partial D_{i,j}}{\partial y}\right). \tag{2}$$

The roughness measure (*Roughness*) is a critical factor affecting ecological and physical processes on the reef (Leon et al, 2015; Dartnell, 2000). It corresponds to the difference between the maximum and minimum depth values over a 3×3 window and is defined for each cell $(i,j) \in [\![1,T]\!] \times [\![1,T]\!]$ as follows:

$$Roughness_{i,j} = \max_{k,l \,\in\, \mathcal{D}_3} (D_{i+k,j+l}) - \min_{k,l \,\in\, \mathcal{D}_3} (D_{i+k,j+l}). \tag{3}$$

The Terrain Ruggedness Index (TRI), is a terrestrial ruggedness measure (Riley et al, 1999) that was adapted to bathymetry data to highlight morphological heterogeneity (Valentine et al, 2004; Rozycka et al, 2017). It is defined as the mean of the absolute differences between the depth value of a cell and the one of its neighboring cells, for each cell $(i,j) \in [\![1,T]\!] \times [\![1,T]\!]$ as follows:

$$TRI_{i,j} = \frac{\displaystyle\sum_{k,l \,\in\, \mathcal{D}_3} |D_{i+k,j+l} - D_{i,j}|}{(3^2 - 1)}. \tag{4}$$

The Vector Ruggedness Measure (VRM) quantifies terrain ruggedness : slope and aspect are decomposed into 3-dimensional vector components using standard vector analysis in a user-specified moving 3×3 window. The vector ruggedness measure is dimensionless because it involves sine and cosine of the slope and aspect measures and its values range from 0 to 1 corresponding to flat regions to rugged ones. Its mathematical definition is omitted to avoid technicalities and details can be found in

(Sappington et al, 2007).
The Curvature position may also be linked to the nature of the seabed. It helps to delimit regions of distinct habitat by identifying boundaries in the character of the terrain ((Wilson et al, 2007)). Bathymetric Position Index (BPI), the marine version of the topographic position index, quantifies where a location on a bathymetric surface is relative to the overall seascape (Mata et al, 2021). It provides an indication of whether any particular pixel forms part of a positive (e.g., crest) or negative (e.g., trough) feature of the surrounding terrain (Lundblad et al, 2006; Wilson et al, 2007). It is calculated using the following formula, for each cell $(i, j) \in [\![1, T]\!] \times [\![1, T]\!]$:

$$BPI_{i,j} = D_{i,j} - \frac{\sum\limits_{k,l \in \mathcal{D}_3} |D_{i+k,j+l} - D_{i,j}|}{(3^2 - 1)}. \tag{5}$$

According to (Evans, 1980), Profile convexity ($Profc$) is the rate of change of *Slope* and Plan convexity ($Planc$) is the rate of change of *Aspect*. Negative values in the $Profc$ indicate the surface is upwardly convex whereas, positive values indicate that the surface is upwardly concave. Positive values in the $Planc$ means the surface is laterally convex and negative values indicate that the surface is laterally concave. Several methods exist for numerical approximations of these metrics, based on a quadratic form representation f of the surface (Florinsky, 1998; Horn, 1981; Evans, 1980; Zevenbergen and Thorne, 1987). Numerical implementation of (Zevenbergen and Thorne, 1987) method's is used in this study ; details can be found in (Florinsky, 1998).
All the computed terrain attributes, depth and geographic coordinates (longitude, latitude) are then stacked to form a multilayer grid of predictors called features or covariates in the sequel Table 1.

Table 1 Terrain attributes computed from the DBM and functions used to do such calculations in R software

Terrain attributes	Reference	Function / R Package
Slope and Aspect		
Slope	(Horn, 1981)	terrain / raster
Aspect	(Horn, 1981)	terrain / raster
Terrain Variability		
Roughness	(Dartnell, 2000)	terrain / raster
Terrain Ruggedness Index (TRI)	(Wilson et al, 2007)	terrain / raster
Vector Ruggedness Measure (VRM)	(Ilich et al, 2023)	VRM / MultiscaleDTM
Curvature and relative position		
Profile Curvature ($Profc$)	(Zevenbergen and Thorne, 1987)	Curvature / spatialEco
Planform Curvature ($Planc$)	(Zevenbergen and Thorne, 1987)	Curvature / spatialEco
Bathymetric Position Index (BPI)	(Ilich et al, 2023)	BPI / MultiscaleDTM

2.2.3 Sampling

To build predictive models for geomorphological typologies using a statistical learning approach, a training dataset is required. This one must contain a finite number of ground truth locations and a set of predictive covariates for geomorphological typologies at these locations. In this section, three alternative clustering based sampling methods are proposed to draw automatically these locations inside a given study area. All these methods use cell's centers of a regular grid and choose among the covariates earlier mentioned depending on the choosen algorithm.

Spatial Coverage Sampling using k-means clustering algorithm (SCS-KMEANS)

The basic idea of Spatial Coverage Sampling (SCS) is to draw uniformly sampling locations over the study area. It has been shown that SCS on a study area can be achieved by k-means clustering algorithm (Hartigan, 1975). This consists in grouping cell's centers of a regular grid on this area using their spatial coordinates as covariates. Note that this regular grid can be the DBM one as long as it does not lead to computional deadlock, otherwise it can be replaced by a raster grid with lower resolution. The final solution of this partition gives the sampling locations and is determined by minimizing a geometric criterion, the mean squared shortest distance between the clusters centroids and the grid cell's centers (Royle and Nychka, 1998; Brus et al, 2006). For the implementation, a k-means algorithm for equal area partitioning is used (Brus, 2019). The *scsKM* function provided in the supplemental materials can be used to achieve this.

Spatial Coverage Sampling using CLARA algorithm (SCS-CLARA)

K-means clustering approach is time and storage consuming, especially for high DBM resolution. In such case, CLARA algorithm approach could be an alternative. The CLARA algorithm is an extension of the Partitioning Around Medoids (PAM) methods (Kaufman and Rousseeuw, 1975) to deal with data containing a large number of objects (more than several thousand observations) in order to reduce computing time and storage problem. Medoids $(M_i)_{i=1}^{k}$ in a PAM, k being the desired number of clusters C_i, are cells that minimize their distance to other cell's centers of the cluster. The CLARA algorithm generates $j \in \mathbf{N}^*$ random samples of size n ($n < T^2$) on individuals, applies a PAM on these samples one after the other, then evaluates the partition quality on each of them by calculating the average global dissimilarity on the complete dataset as follows:

$$\sum_{i=1}^{k} \sum_{x_c \in C_i} \frac{d(x_c, M_i)}{T^2} \tag{6}$$

If this dissimilarity is lower than the previous found one, it considers this solution and its k medoids as the best current solution. The *scsCLARA* function in the supplemental materials can be used to choose ground truth locations.

Complexity-Dependent Sampling using CLARA algorithm (CDS)

Terrain morphology can sometime be marked by accidented region where many typologies are observed closed to each other. This is particularly true with coral reef region where for example sand flat, reef slope, reef flat, etc typologies can be close to each

other. In such situations, homogeneous sampling would lead to miss many typologies unless a lot of points are requested leading to other problems such as important unbalanced typologies distribution among locations (Brus, 2019). To account for such typologies distribution, ground truth locations are agglomerated around zone of higher ground complexity (i.e. complex terrain). Thus the CDS method is introduced in order to take advantage of these covariates available for each DBM resolution. CDS aims to distribute sampling locations around most heterogeneous or complex areas. It uses DBM's cells as individuals and unlike the SCS which uses the spatial coordinates, CDS's covariates are chosen among terrain attributes and depth. Depth and Roughness measures are considered in this study. CDS cannot be computed using the k-means clustering algorithm for high DBM resolutions because DBM cells are mandatory to get covariates. This is why the CLARA algorithm is only used. The *cdCLARA* function in the supplemental materials can be used to this end (Figure 2).

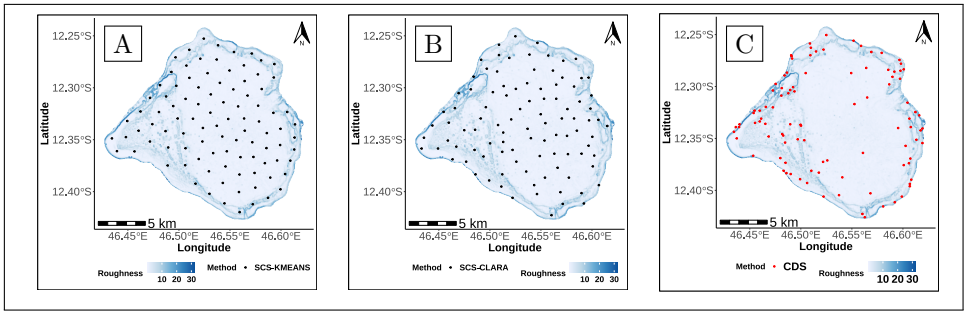


Fig. 2 Example of 100 ground truth locations drawn with the different methods: **A** Spatial Coverage Sampling using K-means clustering algorithm (SCS-KMEANS), **B** Spatial Coverage Sampling using CLARA algorithm (SCS-CLARA) **C** and Complexity-dependant sampling using a CLARA algorithm (CDS). Roughness and Depth are used to guide the complexity-dependant sampling.

2.2.4 Geomorphology mapping using Random Forest algorithm

To map geomorphological typologies over a whole study area, a supervised classification approach is considered. This work was divided in two steps: a first step to train algorithm and a second step to predict typologies based on trained algorithm. For the training part, each location is generated by one of the previously described sampling methods and typologies were attributed using the expert map Figure 1B. These located typologies were the target variable, locations coordinates and the corresponding depth and terrain attributes were used as covariates and both formed the training dataset. In this process, a feature selection scheme random forest based is used to subset the most relevant covariates of the training set. Then a final model is fitted using the selected covariates. After this first step completed, the second step consisted in using the fitted model to predict the most suitable typology over the whole study area.

The generic principle of Random Forest

The tree based Random Forest (RF) algorithm can be used for a classification task

(Breiman, 2001; Biau and Scornet, 2016). Using the Bootstrap AGGregatING (bagging) principle, RF increases the diversity of the trees by making them grow from different randomly drawn (with replacement) training datasets from the original dataset (Breiman, 1996). At each node of each tree, Rf selects a random subset of features and search for the best split for the node. To classify a new case once the forest is completed, the typology having the most votes over all the trees is retained.

Model training

Using a cross-validation scheme with 3 repeats, the training samples are split into 3 folds. To train a RF model, three hyperparameters were tuned: the number of trees ($Ntrees$) of the forest, the number of features used at each node ($Mtry$) and the minimum number of data points at the terminal node of each tree ($nodesize$). Indeed, the $Ntrees$ hyperparameter is not tunable in the classical sense but should be set sufficiently high (Díaz-Uriarte and Alvarez de Andrés, 2006; Oshiro et al, 2012; Probst et al, 2019). The default value of 500 trees is used. The $nodesize$ hyperparameter has been set to 1 for classification task because it generally provides good results (Díaz-Uriarte and Alvarez de Andrés, 2006). The $Mtry$ hyperparameter was tuned among fifteen values of hyperparameters chosen automatically by the function $tuneLength$. A random search optimization strategy which defines a search space as a bounded domain of parameter values and randomly sample points in that domain were used to find the optimal $Mtry$ considering the the resulting Accuracy (Grandini et al, 2020). Models are fitted by repeatedly leaving out one of the folds and performance are determined by predicting on the fold left out. The $train$ function of tthe $caret$ R package were used to this end.

Terrain attributes selection

By selecting the most relevant terrain attributes as covariates, the risk of over fitting can be reduced and the model's generalization ability improved. Variable or feature importance measures are usually used to rank or select variables. Mean Decrease Impurity (MDI) and Mean Decrease Accuracy (MDA) are two well-known random forest variable importance measures (Guyon and Elisseeff, 2020; Breiman, 2001; Biau and Scornet, 2016). In this study, the MDA measure also called permutation importance in Breiman's original random forest is chosen since it seems to exhibit less bias than MDI in presence of correlated features (Strobl et al, 2008; Breiman, 2001). Roughly speaking, the MDA measure consists in shuffling values of a given covariate j in the test data or out-of-bag data (that is data excluded from the boostrap sample used to construct the tree) and then computes the difference between the error on the permuted test set and the original test set. More precisely, for each tree t among the $ntree$ trees of the RF, MDA uses the out-of-bag data to compute a prediction error OOB_t. Then, permuting the values of the j^{th} feature in the out-of-bag data, a prediction error OOB_t^j is computed by using the permuted out-of-bag data. The permutation importance of the feature j is thus defined by $MDA_j = \frac{1}{ntree}\sum_{t=1}^{ntree}(OOB_t - OOB_t^j)$.
To achieve feature selection via the MDA measure, the backward Recursive Feature Elimination (RFE) algorithm (Guyon et al, 2002) implemented in the rfe[4] function of the $caret$ R package is used. This algorithm is based on assessing MDA's importance. MDA is computed by iteratively permuting the values of each input covariate

[4] https://topepo.github.io/caret/recursive-feature-elimination.html

and measuring the resulting drop in prediction accuracy. The feature with the minimum MDA value, representing the least important feature, is systematically removed from the input set. Subsequently, a new RF model is trained using this reduced set of features. This process is continued until the minimal set of input features that yielded optimal Accuracy (Grandini et al, 2020) is obtained. To improve the performance of feature selection with RFE, a repeated 3-fold cross-validation with 3 repeats is used. The *createFolds* function of the *caret* package allowed to split data into training and test sets. This function carries a random sampling within geomorphological typologies in order to balance the classes distributions within the split.
Once the model is trained, typologies can be predicted across the entire study area. Typologies predictions are made using the *predict.train* function of the *caret* R package (Kuhn, 2019).

2.2.5 Performance criteria

To assess model's performance, a *Balanced Accuracy* (BA) metric were calculated using the confusion matrix obtained from each model (Grandini et al, 2020). It consists for each typology (class) k, to calculate a Recall score measuring the ability of a model to find all the positive units for this class as follows:

$$Recall_k = \frac{TP_k}{TP_k + FN_k} \tag{7}$$

True Positives (TP) are observations predicted to belong to the reference class when they really do, and False Negatives (FN) are observations predicted to not belong to the reference class when they really do. BA gives an average measure of this concept using the arithmetic mean of Recall across all classes :

$$BA = \frac{\sum_{k=1}^{K} Recall_k}{K} \tag{8}$$

where K is the total number of class k.
The number of unsampled typologies is used as an indicator of the efficiency of the sampling method to visit all geomorphologic typologies present in the study area.
To assess the consistency of predictions, the map produced at the different DBM resolution are compared to Expert map. For this purpose, the expert map is first dicretized into "pixels" size of 5 m resolution (ie. the smallest resolution used in this study). Then the predicted map grid is disaggregated to match the expert map resolution using *disagRast* function in the supplemental materials. To compare two maps, two metrics are calculated using the confusion matrix between the two: $Match$ and $Balanced\ Match$.
The $Match$ metric is a simple comparison between the two maps. The $Match$ metric corresponding to the proportion of cells identically labelled is calculated as follows:

$$Match = \frac{N_+}{N_+ + N_-} \tag{9}$$

Where N_+ is the total number of cells where typologies match between the two maps, and N_- is the total number of cells that do not match. The *matchRast* function is proposed for the calculation of this metric. However such comparison might hide prediction problems on some small surface typologies, totally dominated by high surface coverage of some typologies. To take account of such fact, the *Balance Match* is a sort of ponderated *Match* removing the surface dominant effect that some typologies might have on others. A second metric denoted *Balanced Match* (BM) inspired by the BA defined in Eq. 8 and using the confusion matrix between a predicted map and the expert map is also calculated. Note that after disaggregating a predicted map, some cells centers around the study site borders may be located outside. These border effects are handled by proportioning BA to the proportion of cells L labelled after the disaggregation of a predicted map. BM is calculated as follows:

$$BM = L \cdot \frac{\sum_{k=1}^{K} Recall_k}{K} \tag{10}$$

The *balmatchRast* function provided in the supplemental materials is used for the BM calculation.

3 Results

3.1 Effect of ground truth sampling methodologies on typologies sampled

Our first methodological assessment was to evaluate the number of missed typologies on the expert map depending on ground truth sampling techniques. Unsurprisingly, the more data points sampled, the greater the chance of sampling all typologies present in the study site, regardless of the sampling method and the resolution of the DBM (Figure 3). Whatever the sampling method and the DBM resolution, small sample size (50 and 100 data points) do not allow to sample all typologies. For these sample sizes, between 1 and 4 typologies are never sampled. It can also be noticed that for 500 data points sampled and more, all typologies are sampled no matter the sampling method and the resolution of the DBM.

3.2 Assessment of model performance based on input data

The *Balanced Accuracy* criteria helps assessing model performance. The larger the sample size, the more precise these measures (ie. smaller standard errors; Figure 4). For small sample sizes (50 and 100 data points), sampling methods give comparable results or even slighty better results for SCS-KMEANS than others sampling methods. In constrast, from 200 points upwards, complexity-dependent sampling tends to give better results than SCS-CLARA and SCS-KMEANS which give comparable results. This result is confirmed as the sample size increases, with a widening gap among sampling methodology means and decreasing standard errors (see Figure S3 and Table S3 for further details).

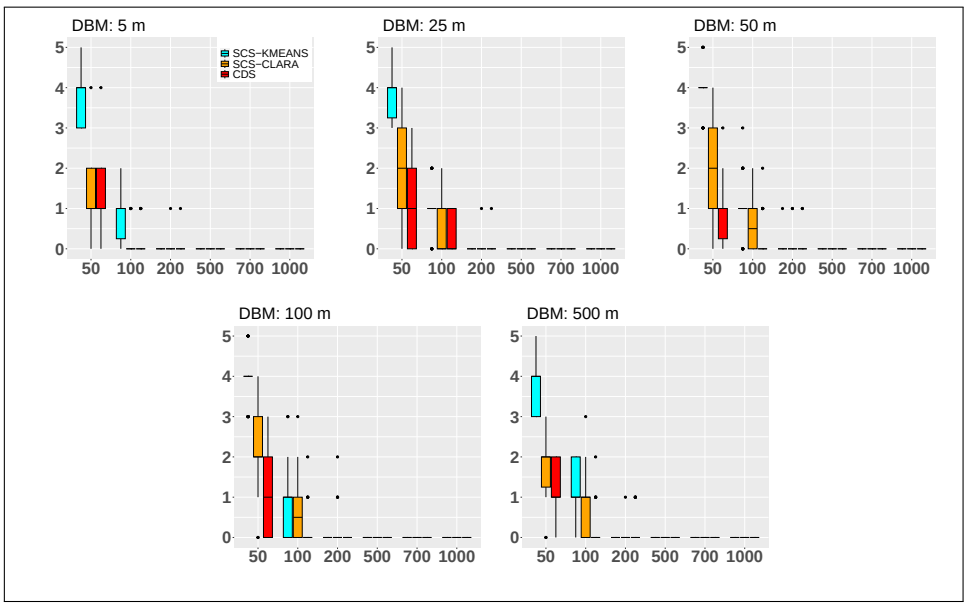


Fig. 3 Boxplot representing the *Number of missing typologies* metric (y-axis) for different sample size (x-axis), different sampling methods (CDS, SCS-CLARA and SCS-KMEANS) and different bathymetric model resolutions (5 m, 25 m, 50 m, 100 m, 500 m). Each sampling conditions were replicated 30 times and represented as default R boxplot settings

3.3 Assessment of selected terrain attributes

A set of features is selected during the feature selection step of the modeling process and general outputs are summarized. Only results for 100, 200, 500 data points sample sizes generated at 100 m DBM resolution using SCS-CLARA method and CDS method are shown her because they are representative of all results (Figure 5). Geographic coordinates (*Latitude* and *Longitude*) and *Depth* are almost always selected regardless the DBM resolution and the sampling method. Terrain variability attributes groups (*Roughness*, VRM and TRI) that are selected a little more than half of the time.

3.4 Evaluation of the quality of produced maps

The *Match* criteria evaluates the consistency of predictions according to the expert map. As previously seen with the *Balanced Accuracy* standard errors values, *Match* errors are non negligible for small sample sizes and decrease when DBM resolutions increase (Figure 6; Figure S3 and Table S4 for further details). In addition, lower *Match* values are recorded at 5 m and 500 m DBM resolutions and higher *Match* values were seen for intermediate DBM resolutions (25 m, 50 m, 100 m) regardless the sample size.

When the sample size and the DBM resolution increase, *Balanced Match* values become more accurate as shown by a decrease of standard errors (Figure 7; Figure S3 and Table S5 for further details). CDS method gives better results than

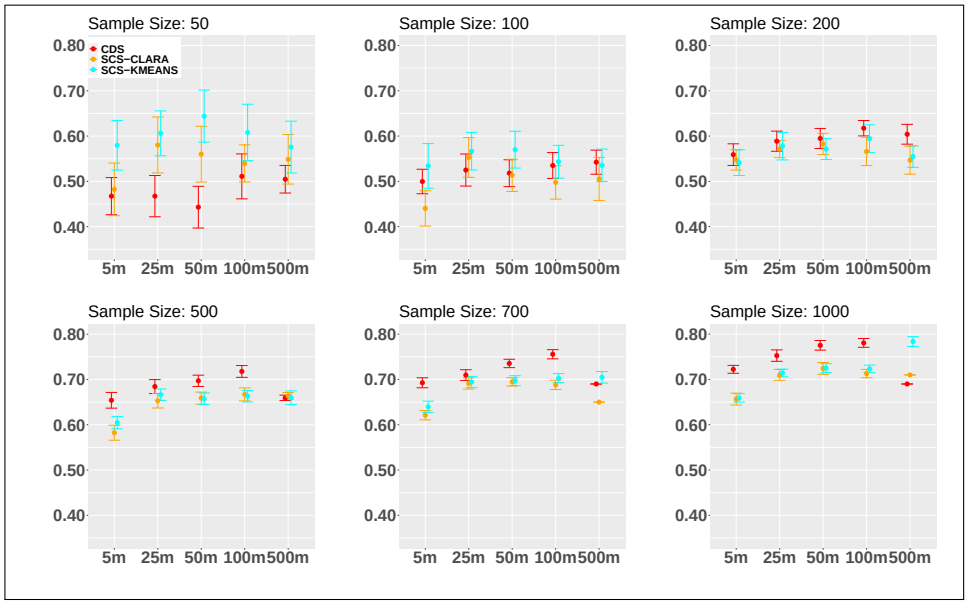


Fig. 4 Mean (+/- standard error) over 30 draws of the *Balanced Accuracy* metric (y-axis) at different DBM resolutions (x-axis) for different sample size and different sampling methods (CDS, SCS-CLARA and SCS-KMEANS)

others spatial coverage sampling methods for any sample size and any DBM resolution. SCS-CLARA and SCS-KMEANS give comparable results according this criteria. Comparing results among DBM resolutions, best results are achieved with 50 m and 100 m DBM resolutions.

In Figure 8 two maps generated with data points sampled using CDS methodology are restituted: (A) 50 m DBM resolution using 1000 data points and (B) 100 m DBM resolution using 200 data points. All typologies are sampled and predicted in both cases. Results show that performance criteria measured are better for the (A') case than the (B') case (see (C)). Indeed, some small surface typologies like Drowned subtidal reef flat and inner slope are better predicted and predictions errors on typologies transition areas less important for (A) than for (B).

4 Discussion

This methodology enabled us to sucessfully reconstruct an expert geomorphological map. According to the *Number of missing typologies* metric, 200 data points are enough to sample almost all the typologies. These locations can be choosen using CDS which performs slightly better than spatial coverage sampling methods (SCS-KMEANS and SCS-CLARA). Results between the different DBM resolutions are comparable for small sample sizes considering the *Balanced Accuracy* metric. But, for 200 data points and more, 100 m DBM resolution gives clearly better results than others resolutions. The *Match* metric supports this finding where 50 m and 100 m DBM

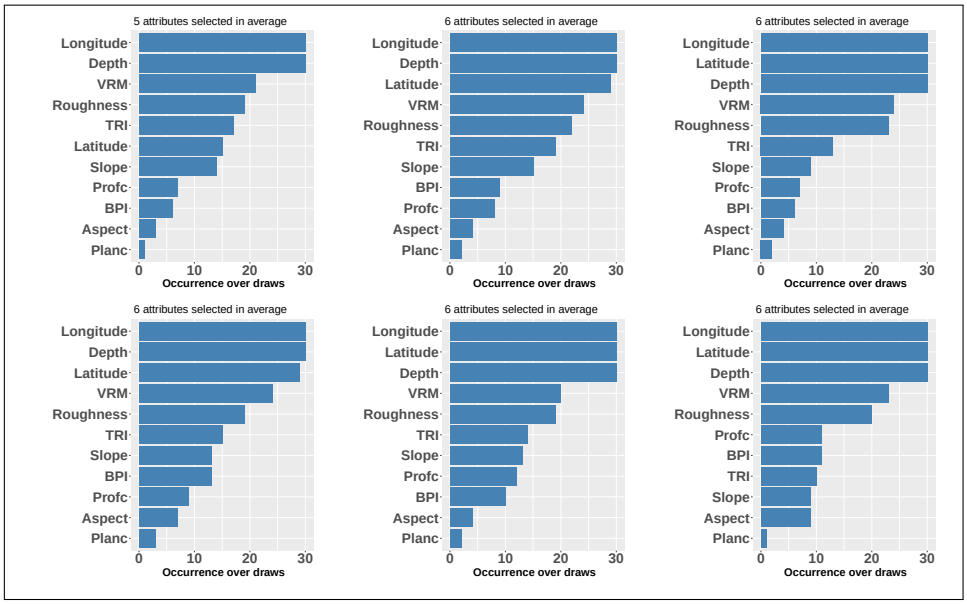


Fig. 5 Ordered occurence over 30 draws of the selected terrain attributes by the RFE algorithm. Here an example with 100, 200, 500 data points sample sizes (from left to right) generated at 100 m DBM resolution using SCS-CLARA method (top) and CDS method (bottom). The average number of attributes selected is noted above each graph (cf. Terrain attribute section for their definition)

resolutions gives much better results than others DBM resolutions whatever the number of data points. Considering the *Balanced Match* criteria which was introduced to contrast the *Match* criteria taking into account typologies surface imbalances, 50 m DBM resolution gives a slight better performance than the 100 m DBM resolution, but considering the sampling effort, it would be preferable to use the 100 m DBM. Indeed, while the sampling effort is multiplied by 5 from (B) to (A), the performance recorded is just a little bit better. We have also seen that the precision of the maps produced is sensitive to the resolution of the DBM, the number and locations of the ground truth selection. Thus, reproductible methodologies with associated codes to evaluate their qualities are proposed. In this section, choices on data sampling to generate such maps are discussed. Then strong and weak points of the modeling approach and alternative to enhance such work are addressed.

4.1 Optimal data acquisition parameters

Creation of a submarine geomorphological map appears, among other consideration, to be constrained between quantity and quality of initial data and cost to acquire and process them. In the present work two kind of data fall in such compromise: the bathymetric data and the ground truth data points.

For the bathymetric data acquisition, high resolution data require advanced technologies, longer survey durations, important storage and sophisticated tools for manipulation which all contribute to higher costs. The results of this study show that a lower DBM resolution do not necessarily lead to the best geomorphological map.

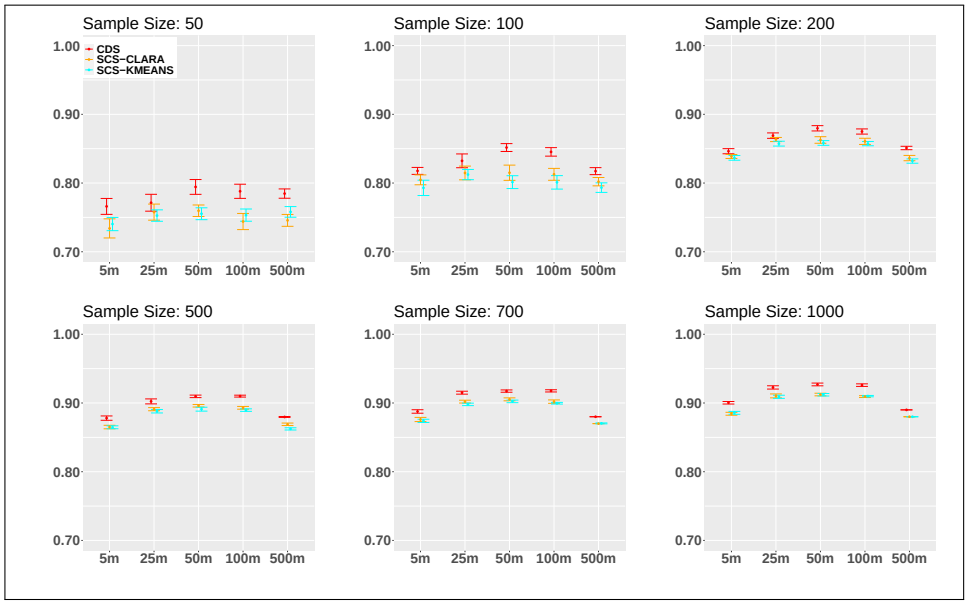


Fig. 6 Mean (+/- standard error) over 30 draws of the *Match* metric (y-axis) at different DBM resolutions (x-axis) for different sample size and different sampling methods (CDS, SCS-CLARA and SCS-KMEANS)

The 100 m DBM resolution considered in this study appears to strike a good balance between detailed geomorphological maps obtained (Figures 4, 6, 7) with low DBM resolutions (5 and 25 m) and large scale geomorphological maps considering (500 m DBM resolution). Indeed, maps at 5 m, 25 m and 50 m DBM resolutions provide a reasonably detailed view of the seafloor and allow the identification of important features but contain a significant amount of noise or small-scale variability that may not be relevant in constructing a geomorphological map. On the other hand, maps of 500 m resolution provide very generalized view of the seafloor representing a significantly coarser scale focusing on major landforms and regional-scale geomorphological patterns but missing important features requiered to build such maps (eg. reef patches, pass or canyons ...). The 100 m resolution was also considered suitable in previous studies for various applications, such as regional planning, environmental assessments, and natural hazard evaluations (Curie et al, 2007; Dong et al, 2019).

Ground truth data acquisition is often acquired by specialists through scubadiving or snorkling (shallow) or using submarine, ROV or drop camera systems allowing to view and characterise seafloor typologies on specific location. In that respect, the number of location is directly correlated to cost of data gathering and therefore need to be balanced. Furthermore, if the number of data point is directly linked to survey cost, the location of sampling on an heterogeneous seafloor structure could be related to data quality (ie. enhanced typology sampling diversity above oversampling of common typologies). To test such compromises, effects number of data points and methodologies to locate them (SCS vs CDS approach) on overall map quality production are

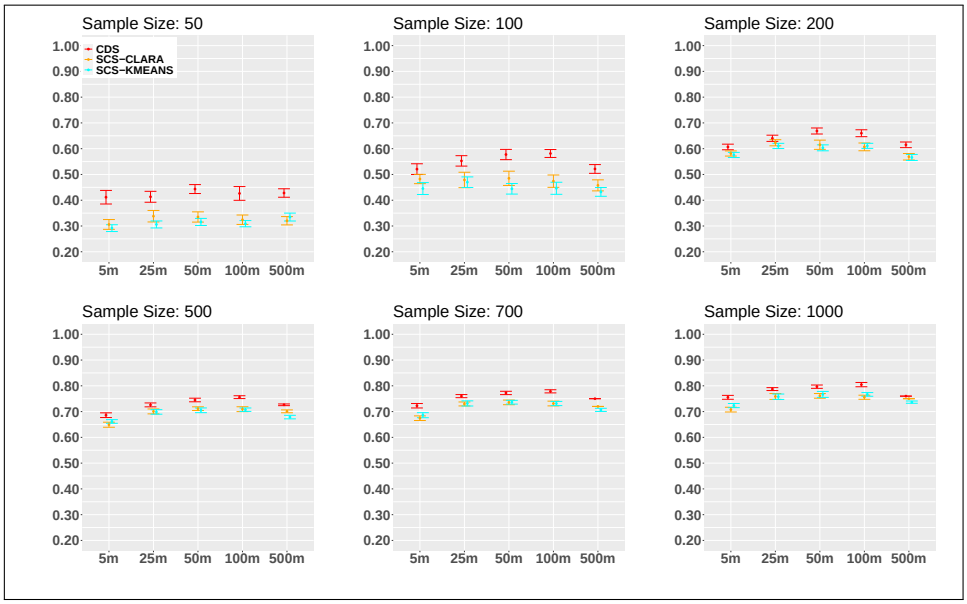
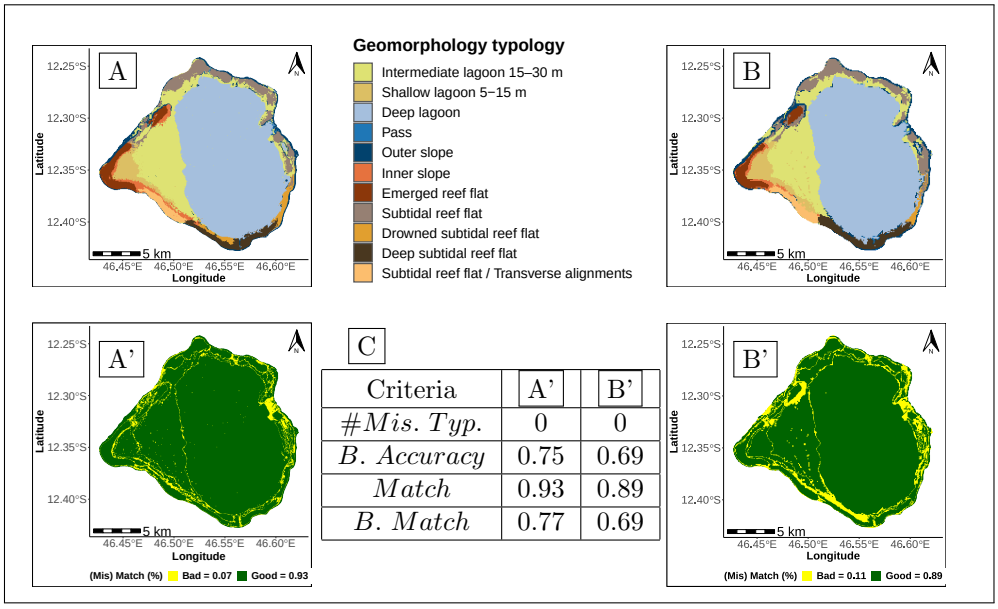


Fig. 7 Mean (+/- standard error) over 30 draws of the *Balanced Match* metric (y-axis) at different DBM resolutions (x-axis) for different sample size and different sampling methods (CDS, SCS-CLARA and SCS-KMEANS).

studied. The choice between homogeneous distribution vs seafloor complexity dependent approach came from the fact that various typologies might agglomerate around location of complex structures (e.g. barrier reef, patch reef, ...). Figure 2 illustrates this point and show that CDS method amplify sampling in area of important typologies diversity. This explains why *Number of missing typologies* criteria results obtained with CDS are better than homogeneous sampling methods (Figure 3). However, such approach has the inconvenient that a bathymetric dataset is required prior planning ground truth sampling campaign. Another point of attention is that attributes used as covariates for CDS algorithms should be carefully selected. This study according to the context of used data (coral reef habitats) focuses on metrics promoted in literature (Adey, 1966; Adey and Macintyre, 1973; Battistini, 1975; Minnery et al, 1985) and on empirical results from this work in the supplemental materials with a supervised classification of geomorphologic typologies based on the terrain attributes presented in Table 1. Seafloor depth and roughness were retained as the most relevant.

Evaluating effect of bathymetric data points density and number and location of ground truth points was done by comparing produced map to an existing expert manually made map. The *Match* criteria were introduced to see how realistic are predictions, regardless the number of sampled typologies. Thus the large geomorphologic units like lagoons (deep, intermediate and shallow), subtidal and deep subtidal reef flats are generally better predicted than the small surface typologies like Pass, Inner slope, Outer slope and Drowned subtidal reef flat typologies. Indeed, such typologies are under-represented in the study area and represent together no more than 10% of the surface. Even evaluating the predictability of such typologies was difficult using



Criteria	A'	B'
#Mis. Typ.	0	0
B. Accuracy	0.75	0.69
Match	0.93	0.89
B. Match	0.77	0.69

Fig. 8 **A** Predicted map with a 50 m DBM and 1000 sampled locations using CDS methodology. **B** Predicted map with a 100 m DBM and 200 sampled locations using CDS methodology. Corresponding match / mismatch maps (A' and B') in comparison to the expert one. **C** Performance criteria measured.

the *Match* metric, hence the use of *Balanced Match* metric pondering typologies by their surface. Using both these metrics, optimal map construction was obtained for 200 ground truth data points obtained using CDS methodology and 100 m DBM resolution (Figure 8).

However, some limitation of the methodology used is to be noted. SCS-CLARA and CDS sampling methods are particularly useful for high dimensional data. The CLARA clustering algorithm can be used on large DBM grid cells data to generate ground truth locations. However, SCS-KMEANS although suitable for spatial coverage sampling, lead to computational deadlock for high DBM resolutions. Furthermore, it can not be used when there is missing data areas in the studied surface. Bathymetric imputation of poorly sampled zones was done using ordinary kriging method that performed much better than others spatial interpolation methods (cf. Table S1, an example on the supplemental materials).

4.2 Modeling choices

Traditional approaches to build geomorphologic maps are often time-consuming, labor-intensive and has a limited coverage and scale. Remote Sensing techniques although allowing wide coverage and high-resolution, require expertise in image interpretation (Gao, 2009; Gilvear and Bryant, 2016). GIS approaches allow for the integration and analysis of diverse data types support data visualization and complex spatial analysis. But they also require specialized software and expertise (Guzzetti et al, 1999; Napieralski and Li, 2007). In addition, interpretation and analysis heavily rely on data

quality. Besides these techniques, semi-automated and automated approaches through machine learning and deep learning are increasingly used to identify complex patterns and features for landform classification, feature detection, or segmentation. Deep learning models require large amounts of labeled training data (e.g., images, point clouds) to effectively learn complex patterns and relationships. Training such models involves adjusting millions of parameters through backpropagation and optimization algorithms (e.g., stochastic gradient descent) and their tuning involves finding the right network architecture and hyperparameters. They are often considered black boxes due to their complex architectures, making it challenging to interpret the features that influence predictions (Li et al, 2020). The statistical learning approach used in this study can handle large volumes of data, automate analysis processes and discover complex patterns and relationships in the data. Interpretability of used models may also be a challenge but less than deep learning ones. In addition, it can work effectively with smaller datasets which was crucial in our objectives.

RF model for geomorphologic units clustering is chosen because RF exhibits complex and non-linear relationships between the features and the dependent variable but also on features among themselves, handling effectively these relationships by using an ensemble of decision trees. It also provide valuable information as feature importance measure helping identify the most informative feature which made it preferable to others clustering techniques in many cases. RF was used for supervised classification problems in many recent studies, compared different algorithms has shown effectiveness of RF based algorithms (Zeraatpisheh et al, 2017; Giaccone et al, 2022). It has shown its robustness specifically when data contain uncertainties and handles high-dimensional data efficiently avoiding overfitting and reducing computational complexity.

The variable selection step in the modeling procedure is crucial in the proposed methodology. Terrain attributes calculation is not specifically time-consuming but the relevance of each of them depends on the data and the geomorphological features that are mapped. This study demonstrated that the terrain attributes, although literature based choosen intially, have not all, a great explanatory power on geomorphologic feature. For each generated sample, the number of features selected is also counted and 6 terrain attributes are generally retained.

5 Conclusion

A statistical learning based approach is proposed to automatically map the geomorphology of a study site using bathymetric data and some ground truth data points. On the one hand, tools to help geomorphologists to plan field campaigns in advance through an optimal DBM resolution and an automated sampling methodology to achieve field verifications are provided. On the other hand, a flexibility in the proposed methodology allowing the usage of terrain attributes as much as desired since the feature selection will help to keep only the most relevant ones is preferred. In addition, statistical and computational tools to compare geomorphological maps produced at different resolutions are provided. The methodology reproducibility is made possible by a set of reusable R scripts.

In the future, an application of the methodology using others data available in others sites is planned. An investigation on others sampling methodology for e.g. which would take into account the presence of non sampling sites inside a study area is also being considered.

6 Supplement information

Additional figures and tables supporting this manuscript can be found after the Reference section.

References

Adey W (1966) Distribution of saxicolous crustose corallines in the northwestern north atlantic. Journal of Phycology 2:49–54. URL https://doi.org/10.1111/j.1529-8817.1966.tb04593.x

Adey W, Macintyre I (1973) Crustose coralline algae: A re-evaluation in the geological sciences. Geological Society of America Bulletin 84(3):883–904. URL https://doi.org/10.1130/0016-7606(1973)84%3C883:CCAARI%3E2.0.CO;2

Ahn S, Sung H, Han H (2023) Classification of the world undersea geomorphic features from gebco 2020 grid data. Journal of the Korean Geographical Society 58(1):36–54

Andréfouët S, Dirberg G (2006) Cartographie et inventaire du système récifalde wallis, futuna et alofi par imagerie satellitaire landsat 7 etm+ et orthophotographies aériennes à haute résolutionspatiale. IRD, Centre de Nouméa et Service de L'Environnement de Wallis et Futuna

Andréfouët S, Muller-Karger F, Robinson J, et al (2004) Global assessment of modern coral reef extent and diversity for regional science and management applications: a view from space. Proceedings of the 10th International Coral Reef Symposium 2:1732–1745

Argyropoulou E, Argialas D, Nomikou P, et al (2016) Automatic identification of submarine landforms using object-based image analysis in the area of north aegan basin. Bulletin of the Geological Society of Greece 50(3):1605–1615. URL https://doi.org/10.12681/bgsg.11880

Arhant Y, Neyt X, Pizurica A (2023) A new deep learning neural network architecture for seafloor characterisation. In The 10th Military Sensing Symposium Proc

Azarafza M, Azarafza M, Akgün PMH.and Atkinson, et al (2023) Deep learning-based landslide susceptibility mapping. Scientific reports 11(1):24112. URL https://doi.org/10.3390/su14031734

Battistini R (1975) Eléments de terminologie récifale indopacifique. Station marine d'Endoume

Behrens T, Schmidt K, MacMillan R, et al (2018) Multi-scale digital soil mapping with deep learning. Scientific reports 8(1):15244. URL https://doi.org/10.1038/s41598-018-33516-64

Biau G, Scornet E (2016) A random forest guided tour. Test 25:197–227. URL https://doi.org/10.48550/arXiv.1511.05741

Bishop M, James L, Shroder Jr J, et al (2012) Geospatial technologies and digital geomorphological mapping: Concepts, issues and research. Geomorphology 137(1):5–26. URL https://doi.org/10.1016/j.geomorph.2011.06.027

Breiman L (1996) Bagging predictors. Machine learning 24:123–140. URL https://doi.org/10.1007/BF00058655

Breiman L (2001) Random forests. Machine learning 45:5–32. URL https://doi.org/10.1023/A:1010933404324

Breyer G, Bartholomä, A.Pesch R (2023) The suitability of machine-learning algorithms for the automatic acoustic seafloor classification of hard substrate habitats in the german bight. Remote Sensing 15(16):4113. URL https://doi.org/10.3390/rs15164113

Browne N, Smithers S, Perry C (2010) Geomorphology and community structure of middle reef, central great barrier reef, australia: an inner-shelf turbid zone reef subject to episodic mortality events. Coral Reefs 29:683–689. URL https://doi.org/10.1007/s00338-010-0640-3

Brus D (2019) Sampling for digital soil mapping: A tutorial supported by r scripts. Geoderma 338:464–480. URL https://doi.org/10.1016/j.geoderma.2018.07.036

Brus D, De Gruijter J, Van Groenigen J (2006) Designing spatial coverage samples using the k-means clustering algorithm. Developments in Soil Science 31:183–192. https://doi.org/https://doi.org/10.1016/S0166-2481(06)31014-8

Copeland A, Edinger E, Devillers R, et al (2013) Marine habitat mapping in support of marine protected area management in a subarctic fjord: Gilbert bay, labrador, canada. Journal of Coastal Conservation 17:225–237. URL https://doi.org/10.1007/s11852-011-0172-1

Cressie N (1988) Spatial prediction and ordinary kriging. Mathematical geology 20:405–421. URL https://doi.org/10.1007/BF00892986

Cui X, Liu H, Fan M, et al (2021) Seafloor habitat mapping using multibeam bathymetric and backscatter intensity multi-features svm classification framework. Applied Acoustics 174:107728. URL http://dx.doi.org/10.1016/j.apacoust.2020.107728

712 Curie F, Gaillard S, Ducharne A, et al (2007) Geomorphological methods to charac-
713 terise wetlands at the scale of the seine watershed. Science of the total environment
714 75(1-3):59–68. URL https://doi.org/10.1016/j.scitotenv.2006.12.013

715 Dartnell P (2000) Applying remote sensing techniques to map seafloor geology/habitat
716 relationships. Masters Thesis, San Francisco State University

717 Dekavalla M, Argialas D (2017) Object-based classification of global undersea topog-
718 raphy and geomorphological features from the srtm30 plus data. Geomorphology
719 288:66–82. URL http://dx.doi.org/10.1016/j.geomorph.2017.03.026

720 Diesing M, Green S, Stephens D, et al (2014) Mapping seabed sediments: Compar-
721 ison of manual, geostatistical, object-based image analysis and machine learning
722 approaches. Continental Shelf Research 84:107–119. URL https://doi.org/10.1016/
723 j.csr.2014.05.004

724 Dong Y, Liu Y, Hu C, et al (2019) Coral reef geomorphology of the spratly islands:
725 A simple method based on time-series of landsat-8 multi-band inundation maps.
726 ISPRS Journal of Photogrammetry and Remote Sensing 157:137–154. URL https:
727 //doi.org/10.1016/j.isprsjprs.2019.09.011

728 Dramis F, Guida D, Cestari A (2011) Nature and aims of geomorphological mapping.
729 Developments in earth surface processes 15:39–73. URL https://doi.org/10.1016/
730 B978-0-444-53446-0.00003-3

731 Díaz-Uriarte R, Alvarez de Andrés S (2006) Gene selection and classification of
732 microarray data using random forest. BMC bioinformatics 7:1–13. URL https:
733 //doi.org/10.1186/1471-2105-7-3

734 Evans I (1980) An integrated system of terrain analysis and slope mapping. Zeitschrift
735 fur Geomorphologic Suppl-Bd 36:274–295

736 Florinsky I (1998) Accuracy of local topographic variables derived from digital eleva-
737 tion models. International Journal of Geographical Information Science 12(1):47–61.
738 URL https://doi.org/10.1080/136588198242003

739 Fukunaga A, Craig B, Kosaki R (2019) Integrating three-dimensional benthic habi-
740 tat characterization techniques into ecological monitoring of coral reefs. Journal of
741 Marine Science and Engineering 7(2). URL https://doi.org/10.3390/jmse7020027

742 Galvez D, Papenmeier S, Sander L, et al (2022) Ensemble mapping as an alternative to
743 baseline seafloor sediment mapping and monitoring. Geo-Marine Letters 42(3):11.
744 URL https://doi.org/10.1007/s00367-022-00734-x

745 Gao J (2009) Bathymetric mapping by means of remote sensing: methods, accuracy
746 and limitations. Progress in Physical Geography 33(1):103–116. URL https://doi.
747 org/10.1177/0309133309105657

Giaccone E, Oriani F, Tonini M, et al (2022) Using data-driven algorithms for semi-automated geomorphological mapping. Stochastic Environmental Research and Risk Assessment 36:2115–2131. URL https://doi.org/10.1007/s00477-021-02062-5

Gilvear D, Bryant R (2016) Analysis of remotely sensed data for fluvial geomorphology and river science. Tools in fluvial geomorphology pp 103–132. URL https://doi.org/10.1002/9781118648551.ch6

Grandini M, Bagli E, Visani G (2020) Metrics for multi-class classification: an overview. arXiv preprint arXiv:200805756 URL https://doi.org/10.48550/arXiv.2008.05756

Guyon I, Elisseeff A (2020) An introduction to variable and feature selection. Journal of machine learning research 3:1157–1182. URL https://doi.org/10.1162/153244303322753616

Guyon I, Weston J, Barnhill S (2002) Gene selection for cancer classification using support vector machines. Machine Learning 46:389–422. URL https://doi.org/10.1023/A:1012487302797

Guzzetti F, Carrara A, Cardinali M, et al (1999) Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, central italy. Geomorphology 31(1-4):181–216. URL https://doi.org/10.1016/S0169-555X(99)00078-1

Hartigan J (1975) Clustering algorithms. John Wiley & Sons, Inc

Horn B (1981) Hill shading and the reflectance map. Proceedings of the IEEE 69(1):14–47. URL https://doi.org/10.1109/PROC.1981.11918

Hugenholtz C, Whitehead K, Brown O, et al (2013) Geomorphological mapping with a small unmanned aircraft system (suas): Feature detection and accuracy assessment of a photogrammetrically-derived digital terrain model. Geomorphology 194:16–24. URL https://doi.org/10.1016/j.geomorph.2013.03.023

Ilich A, Misiuk B, Lecours V, et al (2023) Multiscaledtm: An open-source r package for multiscale geomorphometric analysis. Transactions in GIS 4:1164–1204. URL https://doi.org/10.1111/tgis.13067

Janowski L, Wroblewski R, Dworniczak J, et al (2021) Offshore benthic habitat mapping based on object-based image analysis and geomorphometric approach. a case study from the slupsk bank, southern baltic sea. Science of the Total Environment 801:149712. URL https://doi.org/10.1016/j.scitotenv.2021.149712

Janowski RL.and Wroblewski, Rucinska M, Kubowicz-Grajewska A, et al (2022) Automatic classification and mapping of the seabed using airborne lidar bathymetry. Engineering Geology 301:106615. URL https://doi.org/10.1016/j.enggeo.2022.

106615

Jasiewicz J, Stepinski T (2013) Geomorphons—a pattern recognition approach to classification and mapping of landforms. Geomorphology 182:147–156. URL https://doi.org/10.1016/j.geomorph.2012.11.005

Kaufman L, Rousseeuw P (1975) Partitioning around medoids (program pam). Finding groups in data: an introduction to cluster analysis 344:68–125

Kienholz H (1978) Maps of geomorphology and natural hazards of grindelwald, switzerland: Scale 1: 10,000. Arctic and Alpine Research 10(2):169–184

Koop L, Snellen M, Simons D (2021) An object-based image analysis approach using bathymetry and bathymetric derivatives to classify the seafloor. Geosciences 11:45. URL https://doi.org/10.3390/geosciences11020045

Kuhn M (2019) caret: Classification and regression training. URL https://CRAN.R-project.org/package=caret, r package, version 6.0-92. Accessed: 2023-03-28

Lacharité M, Brown C, Gazzola V (2018) Multisource multibeam backscatter data: developing a strategy for the production of benthic habitat maps using semi-automated seafloor classification methods. Mar Geophys Res 39:307–322. URL https://doi.org/10.1007/s11001-017-9331-6

Lecours V, Dolan M, Micallef A, et al (2016) A review of marine geomorphometry, the quantitative study of the seafloor. Hydrology and Earth System Sciences 20(8):3207–3244. URL https://doi.org/10.5194/hess-20-3207-2016

Leon J, Roelfsema C, Saunders M, et al (2015) Measuring coral reef terrain roughness using "structure-from-motion" close-range photogrammetry. Geomorphology 242:21–28. URL https://doi.org/10.1016/j.geomorph.2015.01.030

Li S, Xiong L, Tang G, et al (2020) Deep learning-based approach for landform classification from integrated data sources of digital elevation model and imagery. Geomorphology 354:107045. URL https://doi.org/10.1016/j.geomorph.2020.107045

Locker S, Armstrong R, Battista T, et al (2010) Geomorphology of mesophotic coral ecosystems: current perspectives on morphology, distribution, and mapping strategies. Coral Reefs 29:329–345. URL https://doi.org/10.1007/s00338-010-0613-6

Lucieer V, Lucieer A (2009) Fuzzy clustering for seafloor classification. Marine Geology 264(3-4):230–241. URL https://doi.org/10.1016/j.margeo.2009.06.006

Lundblad E, Wright D, Miller J, et al (2006) A benthic terrain classification scheme for american samoa. Marine Geodesy 29(2):89–111. https://doi.org/https://doi.org/10.1080/01490410600738021

Maschmeyer C, White S, Dreyer B, et al (2019) High-silica lava morphology at ocean spreading ridges: Machine-learning seafloor classification at alarcon rise. Geosciences 9(6):245. URL https://doi.org/10.3390/geosciences9060245

Masetti G, Mayer L, Ward L (2018) A bathymetry-and reflectivity-based approach for seafloor segmentation. Geosciences 8(1):14. URL https://doi.org/10.3390/geosciences8010014

Mata D, Úbeda J, Fernández-Sánchez A (2021) Modelling of the reef benthic habitat distribution within the cabrera national park (western mediterranean sea). Annals of GIS 27(3):285–298. URL https://doi.org/10.1080/19475683.2021.1936169

Van der Meij W, Meijles E, Marcos D, et al (2022) Comparing geomorphological maps made manually and by deep learning. Earth Surface Processes and Landforms 47(4):1089–1107. URL https://doi.org/10.1002/esp.5305

Minnery G, Rezak R, Bright T (1985) Depth zonation and growth form of crustose coralline algae: flower garden banks, northwestern gulf of mexico. Paleoalgology: Contemporary research and applications Berlin, Heidelberg: Springer p 237–246. URL https://doi.org/10.1007/978-3-642-70355-3_18

Minár J, Evans I (2008) Elementary forms for land surface segmentation: The theoretical basis of terrain analysis and geomorphological mapping. Geomorphology 95(3-4):236–259. URL https://doi.org/10.1016/j.geomorph.2007.06.003

Misiuk B, Brown C (2023) Improved environmental mapping and validation using bagging models with spatially clustered data. Ecological Informatics 77:102181. URL https://doi.org/10.1016/j.ecoinf.2023.102181

Misiuk B, Diesing M, Aitken A, et al (2021) A spatially explicit comparison of quantitative and categorical modelling approaches for mapping seabed sediments using random forest. Annals of GIS 9(6):254. URL https://doi.org/10.3390/geosciences9060254

Napieralski JJ. Harbor, Li Y (2007) Glacial geomorphology and geographic information systems. Earth-Science Reviews 85(1-2):1–22. URL https://doi.org/10.1016/j.earscirev.2007.06.003

Novaczek E, Devillers R, Edinger E (2019) Generating higher resolution regional seafloor maps from crowd-sourced bathymetry. Plos one 14(6):e0216792. URL https://doi.org/10.1371/journal.pone.0216792

Oshiro T, Perez P, Baranauskas J (2012) How many trees in a random forest? Machine Learning and Data Mining in Pattern Recognition MLDM 2012 Lecture Notes in Computer Science(), Springer, Berlin, Heidelberg 7376. URL https://doi.org/10.1007/978-3-642-31537-4_13

Otto JC, Smith M (2013) Geomorphological mapping, vol Section 2.6, British Society for Geomorphology, chap 2, pp 1–10

Otto JC, Prasicek G, Blöthe J, et al (2018) Gis applications in geomorphology. In: Comprehensive geographic information systems. Elsevier, p 81–111, URL 10.1016/B978-0-12-409548-9.10029-6

Pandian P, Ruscoe J, Shields M, et al (2009) Seabed habitat mapping techniques: an overview of the performance of various systems. Mediterranean Marine Science 10(2):29–44. URL https://doi.org/10.12681/mms.107

Pavlopoulos K, Evelpidou N, Vassilopoulos A (2009) Mapping geomorphological environments. Springer Science & Business Media URL https://doi.org/10.1007/978-3-642-01950-0

Probst P, Wright M, Boulesteix A (2019) Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: data mining and knowledge discovery 9(3):e1301. URL https://doi.org/10.1002/widm.1301

Riley S, DeGloria S, Elliot R (1999) Index that quantifies topographic heterogeneity. intermountain Journal of sciences 5(1-4):23–27

Roberts D (2015) Spatially balanced subsampling in r (retaining maximum sample size). https://davidrroberts.wordpress.com/2015/09/25/spatial-buffering-of-points-in-r-while-retaining-maximum-sample-size/

Roos D, Dupont P, Gaboriau M, et al (2017) Projet epicure : Etude des peuplements ichtyologiques et des communautés récifales à partir d'indicateurs spatiaux et de l'approche fonctionnelle, des bancs du geyser, de la zélée et de l'iris. URL https://doi.org/10.13155/54549

Royle J, Nychka D (1998) An algorithm for the construction of spatial coverage designs with implementation in splus. Computers & Geosciences 24(5):479–488

Rozycka M, Migon P, Michniewicz A (2017) Topographic wetness index and terrain ruggedness index in geomorphic characterisation of landslide terrains, on examples from the sudetes, sw poland. Zeitschrift für geomorphologie, supplementary issues 61(2):61–80

Sappington J, Longshore K, Thompson D (2007) Quantifying landscape ruggedness for animal habitat analysis: A case study using bighorn sheep in the mojave desert. Journal of Wildlife Management 71(5):1419–1426. URL https://doi.org/10.2193/2005-723

Schmidt J, Hewitt A (2004) Fuzzy land element classification from dtms based on geometry and terrain position. Geoderma 121(3-4):243–256. URL https://doi.org/10.1016/j.geoderma.2003.10.008

Siart C, Bubenzer O, Eitel B (2009) Combining digital elevation data (srtm/aster), high resolution satellite imagery (quickbird) and gis for geomorphological mapping: A multi-component case study on mediterranean karst in central crete. Geomorphology 112(1-2):106–121. URL https://doi.org/10.1016/j.geomorph.2009.05.010

Siqueira R, Veloso G, Fernandes-Filho E, et al (2022) Evaluation of machine learning algorithms to classify and map landforms in antarctica. Earth Surface Processes and Landforms 47(2):367–382. URL https://doi.org/10.1002/esp.5253

Sklar E, Bushuev E, Misiuk B, et al (2024) Seafloor morphology and substrate mapping in the gulf of st lawrence, canada, using machine learning approaches. Frontiers in Marine Science 11:1306396. URL https://doi.org/10.3389/fmars.2024.1306396

Sowers D, Masetti G, Mayer L, et al (2020) Standardized geomorphic classification of seafloor within the united states atlantic canyons and continental margin. Frontiers in Marine Science 7:9. URL https://doi.org/10.3389/fmars.2020.00009

Stepinski T, Ghosh S, Vilalta R (2007) Machine learning for automatic mapping of planetary surfaces. Aquatic Conservation: Marine and Freshwater Ecosystems 30(4):846–859. URL http://dx.doi.org/10.13140/2.1.1518.9445

Sterne T, Retchless D, Allee R, et al (2020) Predictive modelling of mesophotic habitats in the north-western gulf of mexico. Proceedings of the National Conference on Artificial Intelligence 22(2):1807. URL https://doi.org/10.1002/aqc.3281

Strobl C, Boulesteix A, Kneib T, et al (2008) Conditional variable importance for random forests. BMC bioinformatics 9:1–11. URL https://doi.org/10.1002/aqc.3281

Summers G, Lim A, Wheeler A (2021) A scalable, supervised classification of seabed sediment waves using an object-based image analysis approach. Remote Sensing 13(12):2317. URL https://doi.org/10.3390/rs13122317

Valentine P, Fuller S, Scully L (2004) Terrain ruggedness analysis and distribution of boulder ridges in the stellwagen bank national marine sanctuary region (poster). Galway, Ireland: 5th International Symposium on Marine Geological and Biological Habitat Mapping (GeoHAB)

Wabnitz C, Andréfouët S, Torres-Pulliza D, et al (2008) Regional-scale seagrass habitat mapping in the wider caribbean region using landsat sensors: Applications to conservation and ecology. Remote Sensing of Environment 112(8):3455–3467. URL https://doi.org/10.1016/j.rse.2008.01.020

Wilson M, O'CONNELL B, Brown C, et al (2007) Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope. Marine Geodesy 30:3–35. https://doi.org/10.1080/01490410701295962, URL https://doi.org/10.1080/01490410701295962

Wynn R, Huvenne V, Le Bas T, et al (2014) Autonomous underwater vehicles (auvs): Their past, present and future contributions to the advancement of marine geoscience. Marine geology 352:451–468. URL https://doi.org/10.1016/j.margeo.2014.03.012

Zeraatpisheh M, Ayoubi S, Jafari A, et al (2017) Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in iran. Geomorphology 285:186–204. URL https://doi.org/10.1016/j.geomorph.2017.02.015

Zevenbergen L, Thorne C (1987) Quantitative analysis of land surface topography. Earth Surface Processes and Landforms 12:47–56. URL https://doi.org/10.1002/esp.3290120107

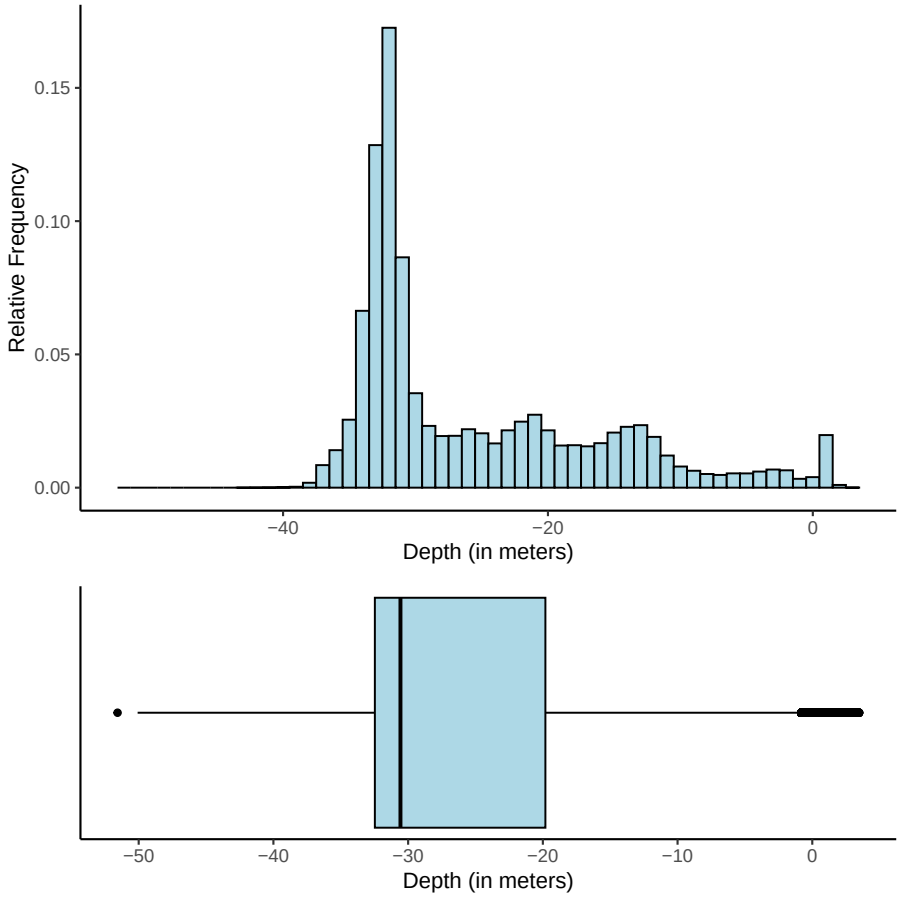


Fig. S1 Distribution of the raw bathymetric data collected on Geyser.

Table S1 Bathymetry imputation of empty grid cells of a 50 m DBM resolution. The Root Mean Squared Error (RMSE) and its Standard Deviation (SD) were calculated using 5-folds cross-validation scheme for five (05) spatial interpolation models : Inverse Distance Weighted (IDW), Nearest Neighboor (NN), Ordinary Kriging (OK), Universal Kriging (UK), Spatial GAM (GAM)

Method	RMSE	SD (RMSE)
IDW	0.80	0.02
NN	0.82	0.02
OK	0.69	0.01
UK	0.71	0.01
GAM	1.34	0.01

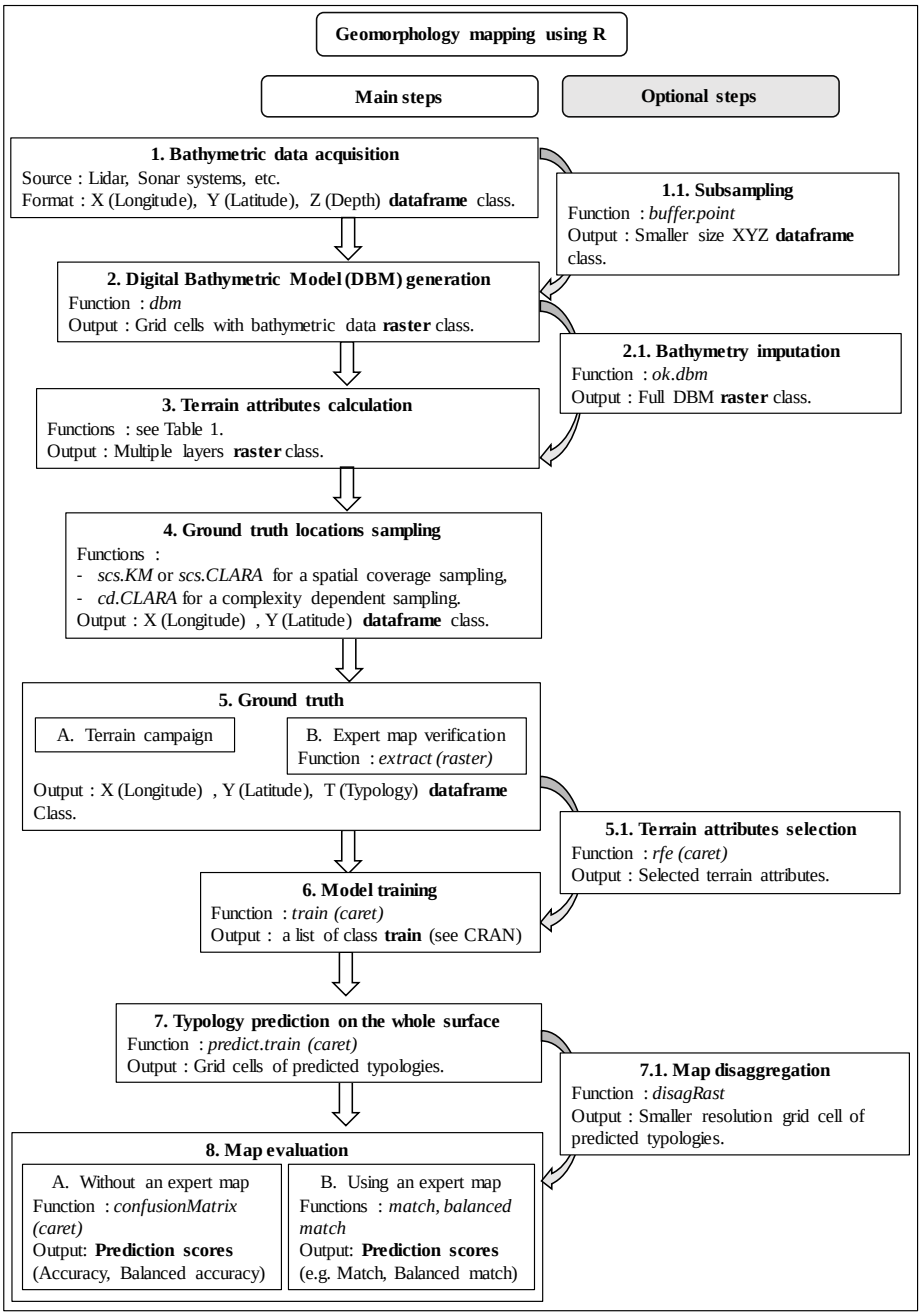


Fig. S2 Statistical learning approach to build geomorphological maps using bathymetric data and typology field verification samples.

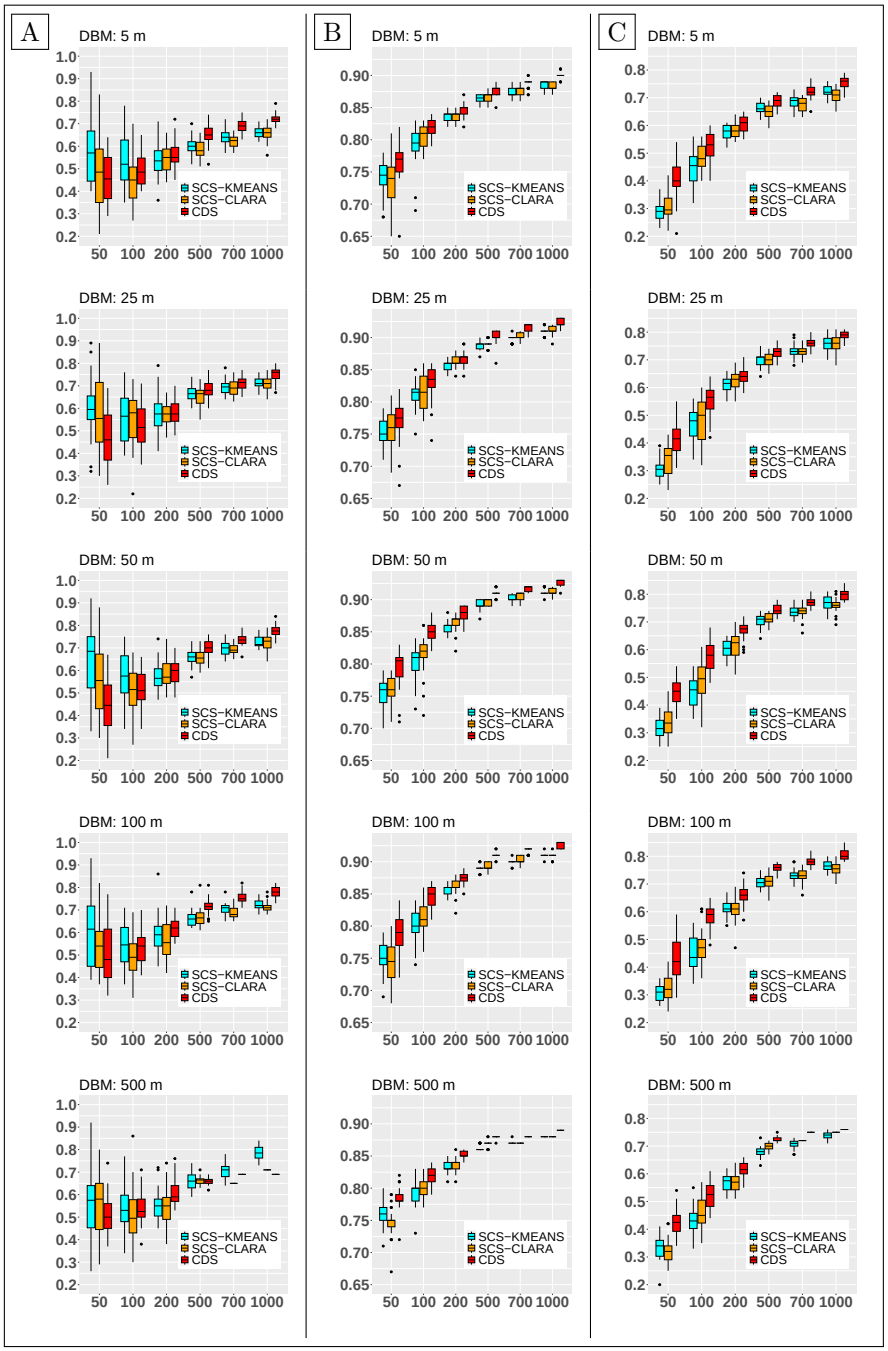


Fig. S3 Boxplot of *Balanced Accuracy* (A), *Match* (B) and *Balanced Match* (C) metrics (y-axis) for different sample size (x-axis), different sampling methods (CDS, SCS-CLARA and SCS-KMEANS) and different bathymetric model resolutions (5 m, 25 m, 50 m, 100 m, 500 m). Each sampling conditions were replicated 30 times and represented as default R boxplot settings.

Table S2 Mean (standard deviation) over 30 draws of the *Number of missing typologies* metric (y-axis) at different DBM resolutions (x-axis) for different sample size and different sampling methods (CDS, SCS-CLARA and SCS-KMEANS)

DBM: 5 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	3.867 (0.681)	0.967 (0.718)	0 (0)	0 (0)	0 (0)	0 (0)
SCS-CLARA	5.033 (0.964)	2.133 (0.937)	0.633 (0.615)	0 (0)	0 (0)	0 (0)
CDS	3.067 (0.944)	0.967 (0.964)	0.167 (0.379)	0 (0)	0 (0)	0 (0)

DBM: 25 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	6 (0.643)	3.1 (0.885)	0.533 (0.629)	0 (0)	0 (0)	0 (0)
SCS-CLARA	5.233 (0.898)	2.367 (0.964)	0.567 (0.626)	0 (0)	0 (0)	0 (0)
CDS	3.3 (1.179)	1.467 (0.9)	0.3 (0.466)	0 (0)	0 (0)	0 (0)

DBM: 50 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	5.933 (0.74)	3.167 (0.791)	0.6 (0.724)	0 (0)	0 (0)	0 (0)
SCS-CLARA	5.067 (0.868)	2.1 (1.125)	0.567 (0.679)	0 (0)	0 (0)	0 (0)
CDS	2.833 (1.053)	0.933 (0.828)	0.233 (0.43)	0 (0)	0 (0)	0 (0)

DBM: 100 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	5.867 (0.681)	3.033 (0.85)	0.567 (0.728)	0 (0)	0 (0)	0 (0)
SCS-CLARA	5 (0.947)	2.2 (0.805)	0.7 (0.702)	0.033 (0.183)	0 (0)	0 (0)
CDS	2.867 (1.252)	0.2 (0.407)	0.133 (0.346)	0 (0)	0 (0)	0 (0)

DBM: 500 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	5.467 (0.73)	2.733 (0.785)	0.467 (0.507)	0 (0)	0 (0)	0 (0)
SCS-CLARA	4.8 (1.031)	2.367 (1.129)	0.4 (0.563)	0 (0)	0 (0)	0 (0)
CDS	3.133 (1.042)	1.433 (0.817)	0.233 (0.43)	0 (0)	0 (0)	0 (0)

Table S3 Mean (standard deviation) over 30 draws of the *Balanced Accuracy* metric (y-axis) at different DBM resolutions (x-axis) for different sample size and different sampling methods (CDS, SCS-CLARA and SCS-KMEANS)

DBM: 5 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	0.58 (0.147)	0.534 (0.132)	0.541 (0.076)	0.605 (0.037)	0.64 (0.033)	0.659 (0.025)
SCS-CLARA	0.482 (0.156)	0.44 (0.105)	0.547 (0.06)	0.582 (0.044)	0.621 (0.027)	0.657 (0.035)
CDS	0.467 (0.11)	0.5 (0.072)	0.559 (0.064)	0.654 (0.047)	0.693 (0.03)	0.722 (0.023)

DBM: 25 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	0.606 (0.133)	0.567 (0.111)	0.578 (0.081)	0.666 (0.034)	0.694 (0.033)	0.714 (0.023)
SCS-CLARA	0.58 (0.166)	0.553 (0.118)	0.571 (0.051)	0.653 (0.043)	0.691 (0.034)	0.71 (0.032)
CDS	0.467 (0.122)	0.525 (0.095)	0.589 (0.06)	0.684 (0.041)	0.71 (0.032)	0.753 (0.034)

DBM: 50 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	0.644 (0.154)	0.57 (0.109)	0.571 (0.062)	0.657 (0.034)	0.697 (0.029)	0.725 (0.026)
SCS-CLARA	0.56 (0.165)	0.513 (0.095)	0.583 (0.063)	0.659 (0.036)	0.695 (0.025)	0.724 (0.035)
CDS	0.443 (0.123)	0.518 (0.079)	0.595 (0.059)	0.697 (0.033)	0.735 (0.025)	0.775 (0.028)

DBM: 100 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	0.608 (0.167)	0.543 (0.097)	0.594 (0.081)	0.663 (0.034)	0.703 (0.027)	0.723 (0.022)
SCS-CLARA	0.54 (0.11)	0.498 (0.099)	0.566 (0.084)	0.667 (0.038)	0.688 (0.027)	0.713 (0.025)
CDS	0.511 (0.133)	0.535 (0.076)	0.617 (0.046)	0.718 (0.035)	0.756 (0.027)	0.781 (0.026)

DBM: 500 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	0.576 (0.153)	0.535 (0.096)	0.555 (0.064)	0.66 (0.04)	0.705 (0.034)	0.784 (0.029)
SCS-CLARA	0.549 (0.146)	0.505 (0.127)	0.547 (0.082)	0.664 (0.02)	0.65 (0)	0.71 (0)
CDS	0.505 (0.082)	0.542 (0.071)	0.604 (0.059)	0.659 (0.016)	0.69 (0)	0.69 (0)

Table S4 Mean (standard deviation) over 30 draws of the *Match* metric (y-axis) at different DBM resolutions (x-axis) for different sample size and different sampling methods (CDS, SCS-CLARA and SCS-KMEANS)

DBM: 5 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	0.74 (0.026)	0.793 (0.03)	0.837 (0.009)	0.865 (0.006)	0.874 (0.006)	0.886 (0.006)
SCS-CLARA	0.734 (0.037)	0.805 (0.019)	0.839 (0.008)	0.865 (0.007)	0.876 (0.008)	0.884 (0.006)
CDS	0.766 (0.031)	0.818 (0.014)	0.846 (0.01)	0.878 (0.008)	0.888 (0.007)	0.9 (0.005)

DBM: 25 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	0.753 (0.022)	0.812 (0.02)	0.857 (0.009)	0.888 (0.007)	0.898 (0.005)	0.909 (0.006)
SCS-CLARA	0.758 (0.031)	0.815 (0.027)	0.863 (0.008)	0.891 (0.006)	0.902 (0.006)	0.91 (0.008)
CDS	0.771 (0.033)	0.832 (0.027)	0.869 (0.011)	0.902 (0.01)	0.915 (0.006)	0.923 (0.006)

DBM: 50 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	0.755 (0.023)	0.801 (0.025)	0.858 (0.009)	0.891 (0.008)	0.902 (0.005)	0.912 (0.005)
SCS-CLARA	0.76 (0.023)	0.815 (0.03)	0.863 (0.013)	0.896 (0.005)	0.905 (0.006)	0.912 (0.005)
CDS	0.794 (0.029)	0.852 (0.015)	0.88 (0.01)	0.91 (0.005)	0.917 (0.004)	0.927 (0.005)

DBM: 100 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	0.753 (0.024)	0.801 (0.026)	0.857 (0.008)	0.89 (0.005)	0.9 (0.003)	0.91 (0.003)
SCS-CLARA	0.744 (0.031)	0.813 (0.023)	0.861 (0.012)	0.893 (0.005)	0.902 (0.007)	0.909 (0.004)
CDS	0.788 (0.028)	0.845 (0.017)	0.875 (0.01)	0.91 (0.004)	0.918 (0.004)	0.926 (0.005)

DBM: 500 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	0.758 (0.021)	0.793 (0.019)	0.832 (0.008)	0.862 (0.004)	0.87 (0.002)	0.88 (0)
SCS-CLARA	0.746 (0.023)	0.802 (0.016)	0.836 (0.01)	0.869 (0.005)	0.87 (0)	0.88 (0)
CDS	0.785 (0.018)	0.817 (0.014)	0.851 (0.007)	0.88 (0.002)	0.88 (0)	0.89 (0)

Table S5 Mean (standard deviation) over 30 draws of the *Balanced Match* metric (y-axis) at different DBM resolutions (x-axis) for different sample size and different sampling methods (CDS, SCS-CLARA and SCS-KMEANS)

DBM: 5 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	0.292 (0.035)	0.445 (0.062)	0.576 (0.027)	0.662 (0.019)	0.686 (0.026)	0.723 (0.021)
SCS-CLARA	0.306 (0.051)	0.482 (0.048)	0.581 (0.027)	0.649 (0.026)	0.674 (0.024)	0.708 (0.025)
CDS	0.412 (0.071)	0.521 (0.056)	0.607 (0.029)	0.686 (0.024)	0.723 (0.023)	0.755 (0.02)

DBM: 25 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	0.306 (0.036)	0.47 (0.055)	0.611 (0.027)	0.699 (0.024)	0.732 (0.027)	0.758 (0.028)
SCS-CLARA	0.338 (0.06)	0.479 (0.08)	0.623 (0.032)	0.701 (0.025)	0.73 (0.021)	0.759 (0.03)
CDS	0.413 (0.057)	0.553 (0.054)	0.64 (0.032)	0.726 (0.02)	0.76 (0.016)	0.787 (0.015)

DBM: 50 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	0.316 (0.037)	0.444 (0.055)	0.603 (0.031)	0.705 (0.025)	0.736 (0.02)	0.767 (0.031)
SCS-CLARA	0.335 (0.053)	0.485 (0.074)	0.615 (0.048)	0.711 (0.019)	0.736 (0.024)	0.761 (0.024)
CDS	0.443 (0.046)	0.577 (0.053)	0.669 (0.031)	0.745 (0.019)	0.772 (0.017)	0.797 (0.017)

DBM: 100 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	0.309 (0.032)	0.446 (0.063)	0.611 (0.026)	0.707 (0.018)	0.731 (0.022)	0.767 (0.019)
SCS-CLARA	0.324 (0.049)	0.474 (0.064)	0.607 (0.04)	0.709 (0.025)	0.731 (0.025)	0.756 (0.021)
CDS	0.426 (0.071)	0.581 (0.041)	0.66 (0.036)	0.756 (0.015)	0.779 (0.016)	0.805 (0.022)

DBM: 500 m

	Sample Size					
Method	50	100	200	500	700	1000
SCS-KMEANS	0.335 (0.041)	0.432 (0.046)	0.566 (0.031)	0.679 (0.018)	0.707 (0.016)	0.737 (0.012)
SCS-CLARA	0.32 (0.043)	0.458 (0.057)	0.569 (0.034)	0.701 (0.015)	0.72 (0)	0.75 (0)
CDS	0.428 (0.044)	0.521 (0.046)	0.615 (0.029)	0.726 (0.009)	0.75 (0)	0.76 (0)