University of Zurich[UZH]

Tackling geomorphological heterogeneity: a comparison of predictors and modelling approaches to assess the distribution of sedimentary organic carbon in submarine canyons

ESS 511 Master's Thesis

Author
Aline Wildberger
08-741-795

Supervised by
Dr. Sarah Paradis Vilar (sarah.paradis@erdw.ethz.ch)
Prof. Dr. Timothy Ian Eglinton (timothy.eglinton@erdw.ethz.ch)
Prof. Dr. Sara Irina Fabrikant

Faculty representative
Prof. Dr. Sara Irina Fabrikant

30.09.2022
Department of Geography, University of Zurich

Abstract

Continental margins are the primary location of organic carbon (OC) burial in the ocean, and extensive efforts have been made to quantify and map the distribution of OC in continental margins worldwide. However, these global estimates do not account for the high geomorphological heterogeneity in continental margins, such as the influence of submarine canyons as sites of preferential OC deposition.

This thesis compares the accuracy of geostatistical external drift kriging with different machine learning approaches to predict surficial sediment OC content in the extensively studied Nazaré Canyon (Central Portuguese margin) and its adjacent continental margin. Random forests using explicitly spatial covariates performed slightly better (RMSE=0.527 wt. %) than classical random forests (RMSE=0.534 wt. %) and predict the spatial distribution of OC substantially better than external drift kriging (RMSE=0.705 wt. %). Distance to the canyon axis and surface rugosity are amongst the most important predictors of OC content in the canyon margin system and challenge thus factors known to affect the distribution of the latter in surficial sediment.

The Nazaré Canyon contains 46% more OC per unit area (151 $g \cdot m^{-2}$) in comparison to its adjacent continental margin (103 $g \cdot m^{-2}$), proving that canyons are important sites of OC deposition due to their capacity of funnelling large quantities of sediment and OC towards the deep sea. Considering that submarine canyons incise all continental margins, occupying 10% of their global area, and that the present findings suggest substantially higher carbon contents within those features, they should be accounted for in global estimates of OC deposition in continental margins worldwide.

Acknowledgements

Thank you

- … Sarah Paradis, my main supervisor, for helping me in all circumstances and answering all my burning questions. She was a great support throughout his whole thesis and made me learn a lot!
- … Tim Eglinton, who introduced me yet to another ocean and mapping related project
- … Sara Fabrikant, who shared parts of her extensive cartographical knowledge with me, introduced me to new tools and gave me good feedback
- … Andreas Papritz for his great help during the geostatistical modelling of the data and for discussing spatial analysis
- … Aleksandar Sekulić for writing a great R package and giving me input throughout the modelling process
- … Cheng Fu for sharing his machine learning knowledge with me
- … Devin Routh for his input regarding spatial prediction and machine learning and the discussions on mapping related topics
- … Gereon Kaiping and Ross Purves for reviewing my concept and all the input you have provided
- … Daniela Mariño and Mikhail Kanevski for input on geostatistics
- … Eun-Kyeong Kim for the ressources on kriging
- … Myles Lewis for tweaking his R package for serve my purpose
- … Elke Kossel & Matthias Haeckel, Veerle Huvenne, Roberto Martins, Anabela Oliveira, Caroline Slomp, Henko de Stigter and Laurenz Thomsen for their data and/or clarifications regarding the latter
- … Thomas Werschlein for solving all my IT problems within no time and helping me survive the last few weeks of analysis
- … Philipp, for helping me extract a thalweg from the canyon and for any other remote sensing related input
- … Margot for reviewing my thesis last minute and for her great ideas to improve it
- … The S3IT team from UZH for navigating me through their infrastructures
- … Tumasch Reichenbacher for helping me setting up the concept talk and allowing me promptly the access to the S3IT infrastructures
- … The Hydrographic Institue of Portugal for providing their highly resolved local bathymetry
- … All researchers who have publicly made available their sediment data

Table of Contents

1 Introduction

1.1 The continental margin and carbon storage

Sedimentary rocks host the largest proportion (~ 90%) of organic carbon (OC) globally (Hedges & Keil, 1995). If successfully buried, marine surface sediment organic carbon can escape into the long-term carbon cycle and remain there over geological timescales, sequestering carbon away from the atmosphere (Atwood et al., 2020). Marine sediments play therefore a key role as climate regulators.

Although continental margins occupy only 10-20% of the global ocean floor by area, they store up to 90% of the OC being preserved in marine sediments (Hedges & Keil, 1995), and thus are recognized as significant stockholders of carbon in the marine realm (Bianchi et al., 2018) and are the primary location of OC burial in the ocean (Ausín et al., 2021). However, this storage is increasingly threatened due to changing climate, coastal development, mining, hydrocarbon exploration (Cordes et al., 2016) and commercial fishing methods that involve deep-sea bottom-trawling (Sala et al., 2021).

In order to guide informed ocean management decisions, e.g. defining Marine Protected Areas (MPA), precise OC stocks need to be presented to policy makers (Atwood et al., 2020; Diesing et al., 2021; Smeaton, 2021). As continental margins comprise a multitude of different settings (Avelar et al., 2017) their ability to accumulate and bury OC has to be assessed taking into account this diversity. The IPCC (IPCC Working Group 1 et al., 2013) still uses spatially non-explicit estimates for sedimentary carbon reservoirs, based on a simple multiplication of the global extent of open ocean and continental margins multiplied by a respective mean OC content (Emerson & Hedges, 1988). This binary view needs to be replaced by a more nuanced integration of a heterogeneous continental margin into calculations of carbon reservoirs.

Submarine canyons are a prime example for the geomorphological heterogeneity of the seafloor and are widespread features, present on the majority of the continental margins. Over 9000 large canyons cover over 10% of global continental margin area (Harris et al., 2014) displaying a steep and complex topography (Harris & Whiteway, 2011), resulting in diverse current patterns (Xu, 2011) and providing a multitude of habitats (de Leo et al., 2014). Despite their roles as "keystone structures"(Vetter et al., 2010) and the main pathway for sediment transport from the shelf to the deep sea, submarine canyons have not yet been recognized as hotspots for the remineralization or sequestration of OC. Over the last few years, research has focused on sediment transport mechanisms (Allin et al., 2016; Arzola et al., 2008; Puig et al., 2014) and the influence of bottom trawling (Paradis et al., 2017; Payo-Payo et al., 2017), benthic habitat composition (Appah et al., 2020; Huvenne et al., 2012), and the geochemistry of seafloor sediments (García et al., 2008; Kiriakoulakis et al., 2011; Oliveira et al., 2011). But precise techniques to spatially interpolate sedimentary OC in these highly heterogeneous environments have not been employed yet.

The deposition and subsequent preservation of organic carbon (in particulate form (POC)) in marine sediments is governed by a multitude of factors. Marine POC stems from different sources and encounters obstacles before and during its gravitational settling to the sea floor. We can distinguish between allochthonous POC, organic carbon derived from the remains of marine primary producers (algae) and allochthonous POC, matter of terrestrial origin (e.g. vascular plant detritus, black carbon or fossil OC from the erosion of meta-sediments) (Kandasamy & Nath, 2016; Kharbush et al., 2020). Although marine OC production (around 50,000 Tg C yr^{-1}) largely surpasses the input from land to the ocean (around 740 Tg C yr^{-1}), burial efficiencies of the latter are much higher: 14% of all riverine TOC and 10% of aeolian POC are eventually buried compared to 0.8% and 0.03% for OC derived from coastal ocean and open ocean primary productivity, respectively (Kandasamy & Nath, 2016). As POC

is adsorbed to sediment particles, it follows the path the sediment takes. Once deposited as bottom-sediment, it can be further transported by bioturbation or laterally moved upon resuspension of the sediment in the water column (LaRowe et al., 2020). Therefore, factors such as nepheloid layers, strong bottom water currents or mass wasting events, move deposited sediment and can locally increase or decrease carbon content. Various other post-depositional processes can affect the content of total organic carbon (TOC, in this thesis referring to the particulate fraction only) in surficial sediments. For example, decomposition of organic constituents occurs mostly in bioturbated sediments below oxygenated waters (Wakeham & Canuel, 2006).

1.2 Ways to predict and assess the spatial distribution of TOC

The prediction of a variable of interest (TOC in the present case) at unsampled locations based on sparse ground truth data can be tackled in different ways. Early interpolators such as Thiessen Polygons (Thiessen, 1911) and triangulation based on Delaunay (1934) take into account only one or three ground truth points, respectively, for predicting a value and showcase a more or less pronounced step-like appearance. The extension and combination of the former two, the natural neighbour interpolant (Sibson, 1981), can bypass the emergence of abrupt changes within the prediction surface but produces strongly biased results where the data is noisy (Webster & Oliver, 2008). The inverse distance weighting (IDW) method (Shepard, 1968) obeys the popular and only years later formulated law of Waldo Tobler first law of geography stating "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). All these interpolators are local, meaning they calculate the predicted value from neighbourhoods, smaller in extent than the full study area. Global interpolators on the other hand use the entire study area spanning dataset. Trend surface analysis (Krumbein, 1959), a global interpolation technique in that it uses the full study area spanning dataset, is based on polynomial regression of the x and y coordinates of the ground truth locations. The resulting regression equation acts then as the interpolator. As the spatial variation of a variable is often complex, a polynomial of high order would be needed to account for it and this would result in highly unstable matrix equations. Additionally, the autocorrelated structure of the trend residuals violates one assumption of linear regression models, namely the independence of errors (Poole & O'Farrell, 1971).

All of the above-mentioned spatial interpolation techniques are deterministic, which means that they assume the variation in a variable is determined by physical causes. But processes that influence the variation of an environmental variable often vary in a strongly non-linear and chaotic ways. The value of a variable at a location can be seen as the outcome of a random process (i.e., each value is drawn at random from a probability distribution), therefore the spatial variation should be considered as random (Matheron, 1965). Linear kriging, the workhorse of geostatistics, offers a solution to account for the randomness within the variance. It uses the spatial autocorrelation of measurements (near things being more similar than distant things) to its advantage by stratifying the environment to calculate predicted values as weighted linear sums. It can be seen as a sophisticated version of IDW and differs from the later in that its weights are based on the degree of spatial correlation, based on a semivariogram, and not on spearation distance between measurements (Sekulić et al., 2020). In the case of kriging, a model is fitted to the semivariogram, displaying the semivariance with increasing lag distance between sampled points. A covariance matrix based on the estimated variogram is then used to calculate the weights which will be plugged into the kriging predictor. To ensure unbiased estimates, the weights are constrained to sum up to 1 (for the most common kriging procedures like ordinary kriging and its extensions). Additionally, each predicted location has an associated variance

which allows for a measure of uncertainty, e.g. the standard error, a local validity check that is not given for the other methods.

The kriging predictor, also named BLUP (best linear unbiased predictor), can only live up to its full potential when the target variable is normally distributed. Non-normal data can be transformed for the sake of kriging but if the data exhibits non-gaussianity and has been generated by highly non-linear processes, the estimation can be a challenging task.

Recently, machine learning approaches, e.g., k nearest neighbours (kNN, (Altman, 1992)) and random forests (Breiman, 2001), have become increasingly popular for the spatial prediction of continuous and discrete variables (e.g. classification of land use and cover LUC) and are often used instead of or in combination with geostatistics (Kopczewska, 2022). They are not underpinned by rigid statistical assumptions (Erdogan Erten et al., 2022), showcase higher flexibility when it comes to incorporating and combining covariates of different types, and can maneuver complex non-linear relationships (Fouedjio & Klump, 2019).

Although they make no assumption about the underlying spatial distribution of the variable, random forests still assume the data to be i.i.d. (independent and identically distributed, which means spatial autocorrelation is ignored) which is often unrealistic, even more so when there is an uneven sampling density. Another limitation lies in the fact that uncertainty quantification, e.g. the kriging variance in geostatistics, is not given for most machine learning approaches (including the classic random forest approach) and the algorithms do not reproduce data at the sampled locations (Erdogan Erten et al., 2022). Lately, attempts have been made to render the random forest framework more suitable for spatial applications, i.e., by acknowledging spatial autocorrelation. Authors have been adding spatial covariates like distance buffers (Hengl et al., 2018) and the values of and distances to nearest neighbours (Sekulić et al., 2020). Furthermore, a spatial validation procedure of the models has been implemented inter alia by Ploton et. al (2020) and Meyer et al. (2018). The debate about the suitability of machine learning models and spatial validation techniques within the spatial realm is ongoing (Chen et al., 2019; Fouedjio & Klump, 2019; Meyer & Pebesma, 2022; Wadoux et al., 2021).

1.3 Research goals and related questions

The goals of this thesis are to build a spatially highly resolved model of sedimentary organic carbon content of a submarine canyon and its adjacent continental margin and to produce a cartographic representation that emphasizes the potential carbon storage capacity of submarine canyons.

The following research questions will be answered within this framework:

1. How accurately do different geostatistical and machine learning interpolation techniques predict surficial sediment TOC?
2. How influential are different predictor variables in explaining local changes in TOC content within surface sediments of submarine canyons?
3. How does geomorphological heterogeneity affect the spatial distribution of surficial sediment TOC content?
4. How much TOC is sequestered in surficial sediments of the Nazaré submarine canyon in comparison to its adjacent continental margin?

2 Related work

Numerous studies to predict and map sedimentary organic carbon contents within the marine realm have been conducted. These studies predict TOC from global to local scales and differ not only in the chosen prediction approaches but also in the resulting spatial resolution of the predicted surfaces.

Published estimates of global surficial sediment organic carbon stocks vary a lot: they range from 87 for the top 5 cm (Lee et al., 2019) to 3117 Pg for the top 1 m (Atwood et al., 2020). These differences can not be explained solely by the fact that different reference depths were considered, which underlines the importance of developing adequate methods which can better constrain stocks of TOC (Diesing et al., 2021).

Early attempts by Romankevič et al. (1984) and Premuzic et al. (1982) mapped surficial sediment TOC on a global scale but without spatially explicit estimates. They used average TOC contents for bigger sub areas of the world's ocean (basins and continental margins), thus the resulting maps from these large-scale interpolations offer only a low-resolved overview. Distinct patterns within the continental margins are not extractable, but they still capture lower TOC contents in basins (< 0.5 wt. %) and higher contents (> 0.5 wt. %) along the continental margins (Seiter et al., 2004). Some type of kriging has been used to predict TOC in the South Atlantic Ocean (Mollenhauer et al., 2004), the Eastern Arabian shelf (Acharya & Panigrahi, 2016) and for the Global Ocean (Seiter et al., 2004). Surficial sediment TOC has been constrained in the Gulf of Mexico with isopletes (Escobar-Briones & García-Villalobos, 2009) and Neto et al. (2016) took advantage of a confirmed relationship between TOC content of marine sediments and seismic peak amplitude, laterally extrapolating between closely spaced seismic profiles.

The use of spatially exhaustive explanatory variables (e.g., satellite data) to predict TOC contents, whose sampling is more time-consuming and expensive has been gaining ground within studies aiming to predict carbon contents in marine sediments. Lee et al. (2019) used kNN to gain insight into the distribution of seafloor TOC globally, while Atwood et al. (2020) utilized random forest to tackle the same task. Limited to local extents, Markus Diesing and colleagues have predicted organic carbon contents of the North-West European continental shelf (Diesing et al., 2017) and the North Sea and Skagerrak Strait (Diesing et al., 2021) using random forests, and quantile regression forests (QRF), respectively. QRF (Meinshausen, 2006) is an extension of random forests, allowing to derive uncertainty estimates.

Studies focusing on the prediction of continental margin TOC are rare, and those that focus on the distribution of organic carbon within canyons even more so. The only published study that explicitly mapped the distribution of sedimentary TOC within a submarine canyon environment is from Baudin et al. (2017). They used the deterministic natural neighbour interpolant to derive surface TOC values for the Congo deep-sea fan which is fed by the Congo submarine canyon.

Machine learning and/or geostatistical approaches have not yet been used to derive TOC contents within a submarine canyon, although an abundance of auxiliary variables are at our disposal, ready to be used as predictors in machine learning or in kriging approaches that implement the use of covariates other than the geographical coordinates (e.g. kriging with external drift (Delhomme, 1979)) to model a trend in the response variable.

An accurate spatial model for submarine canyons and its adjacent margin would not only help to constrain carbon budgets on continental margins more precisely, but would ultimately help constrain global budgets as well, which is of uppermost importance in present times.

3 Case study: The Nazaré submarine canyon

The Nazaré Canyon (Figure 3.1) is one of three Central Portuguese submarine canyons on the Western Iberian continental margin. The 210 km long feature dissects the margin in an east-west direction from -50 m down to the Iberian abyssal plain at almost -5000 m. Arzola et al. 2008 showed that upper (proximal) sections of the canyon are governed more by erosive processes, whereas lower (distal) sections more by depositional ones. The steep topography in the proximal canyon section results in an instability that can provoke intra-canyon landslides and rock avalanches. During earthquakes, slope failures on the shelf break and around the canyon head release impressive volumes of sediments that get flushed through the canyon and are deposited predominantly in the lower canyon section and up to the abyssal plains.

As major pathways for land derived sediment (and therefore also organic carbon) to the deep ocean, the Nazaré Canyon receives material via different transport mechanisms. In summer there is resuspension of northern mid shelf sediments by internal wave activity and subsequent lateral transport towards the canyon by upwelling currents. Along-shelf transport of eroded sea cliff and beach material to the north and south of the canyon happens all year round. In winter, under downwelling conditions, fine fluvial sediment from rivers to the south of the canyon can reach its head. During winter floods, the nearest fluvial source, the Mondego River, can also input sediment to the shelf (Arzola et al., 2008).

Higher organic carbon contents are associated with fine sediment, as observed in the Nazaré Canyon by Oliveira et al. (2007). Sediments of the Nazaré Canyon have consistently higher organic carbon content than the ones from the close slope (Kiriakoulakis et al., 2011; Masson et al., 2010; Oliveira et al., 2007) and the canyon itself is a preferential site for the accumulation of fine-grained sediments (Schmidt et al., 2001). There is also an indication that tributaries and the canyon head (around -300 m) are permanent depocenters (areas of maximum deposition, (Oliveira et al., 2007)).

The first attempt to constrain TOC accumulation and subsequent burial in the Nazaré Canyon (Masson et al., 2010) was based on an oversimplifying approach, dividing the canyon into four zones of distinct sedimentation rates and mean TOC content based on canyon morphology. This subdivision of space may hinder though derivation of accurate TOC stocks. Using the richness of site-related sediment cores as well as backscatter data, this thesis seeks to constrain carbon accumulation along this morphologically complex canyon (Lastras et al., 2009), building upon recent studies and modelling this heterogeneous environment while circumnavigating sample scarcity (Atwood et al., 2020; Diesing et al., 2021, 2017, 2014; Jerosch, 2013; Mitchell et al., 2021).

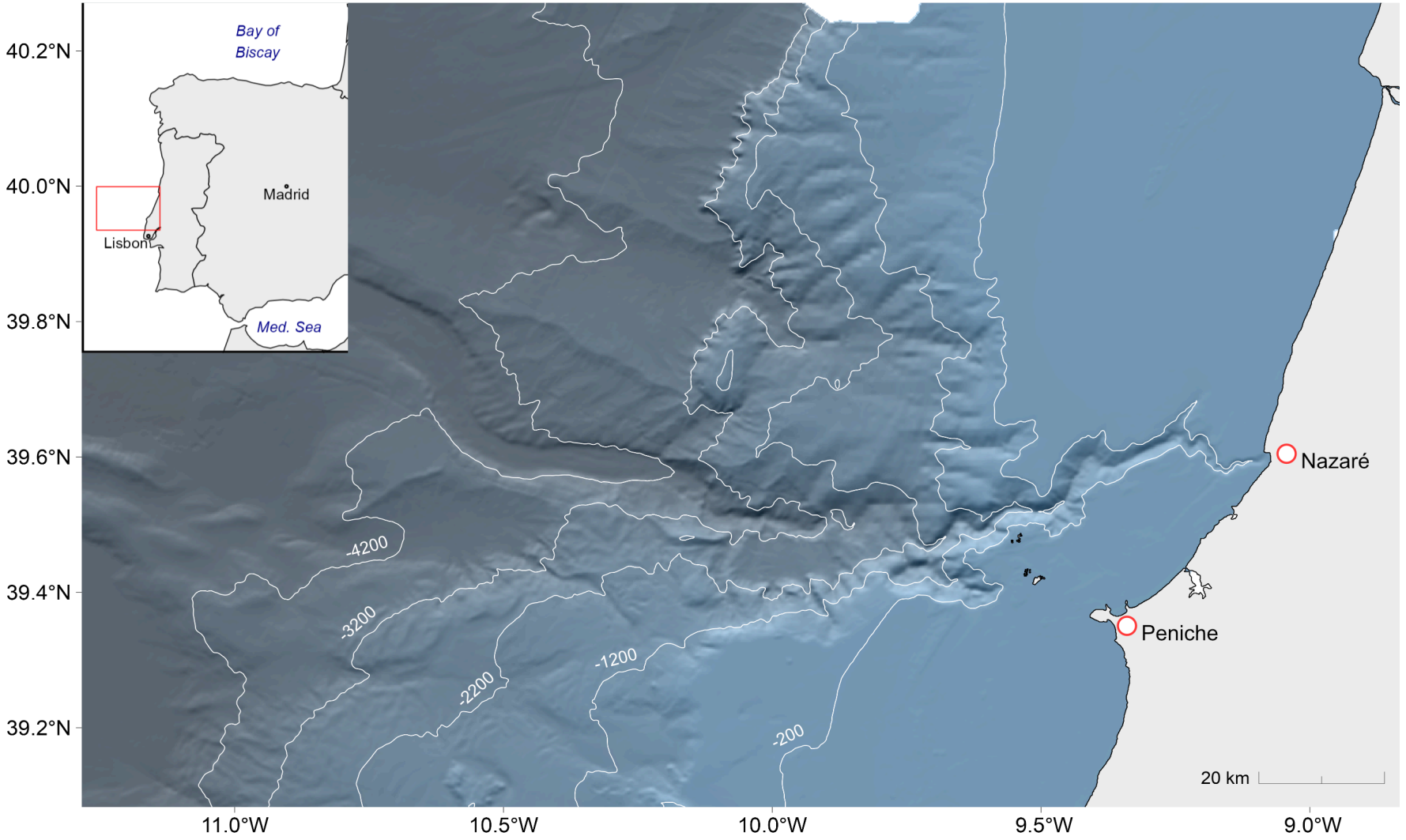

Figure 3.1 Overview of the study site: Isobaths with corresponding depths [m] (white solid lines).

4 Materials and Methods

4.1 Data acquisition and processing

4.1.1 Surficial sediment organic carbon

Data of marine sedimentary total organic carbon content [wt. %] for the chosen study area (delimited by the extent of the greatest common area of the available covariates) was queried from the MOSAIC database (van der Voort et al., 2021) or has been manually extracted from scientific publications.

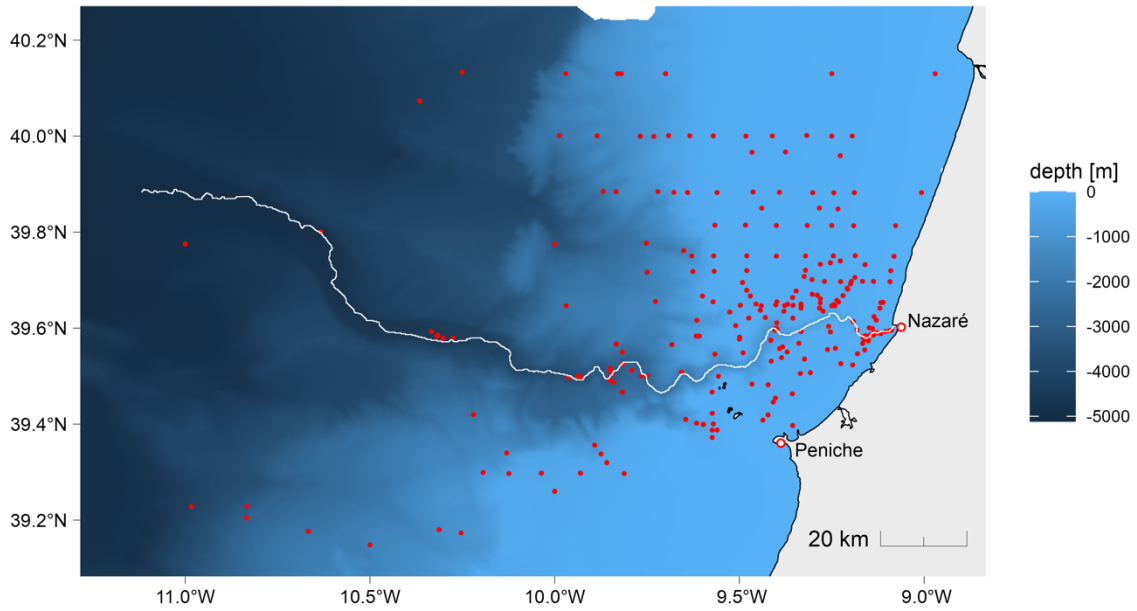

Figure 4.1 Sediment core locations. Red dots represent stations where sediment cores for surface sediment organic carbon were retrieved.

The available stations (Figure 4.1) stem from 16 sampling campaigns (within the period of 1997 to 2011) and have been incorporated into at least 16 scientific publications. Most cores had been retrieved during meteorological spring and autumn (Figure 4.2) and using different sampling devices (Table 1). Figure 4.3 shows that there is no clear dependence of the sampling method on the sediment content at different depths, except higher contents for the top 0.5 cm with the multi corer.

Table 1 Sampling devices used for core retrieval of study samples

sampling method	# of cores
grab corer	157
multi corer	75
box corer	23
push corer	6
piston corer	1
gravity corer	1

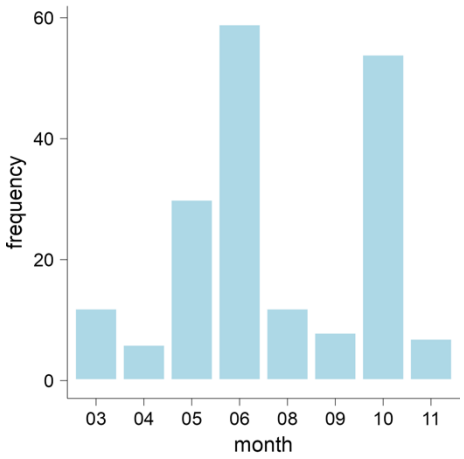

Figure 4.2 Month of core retrieval

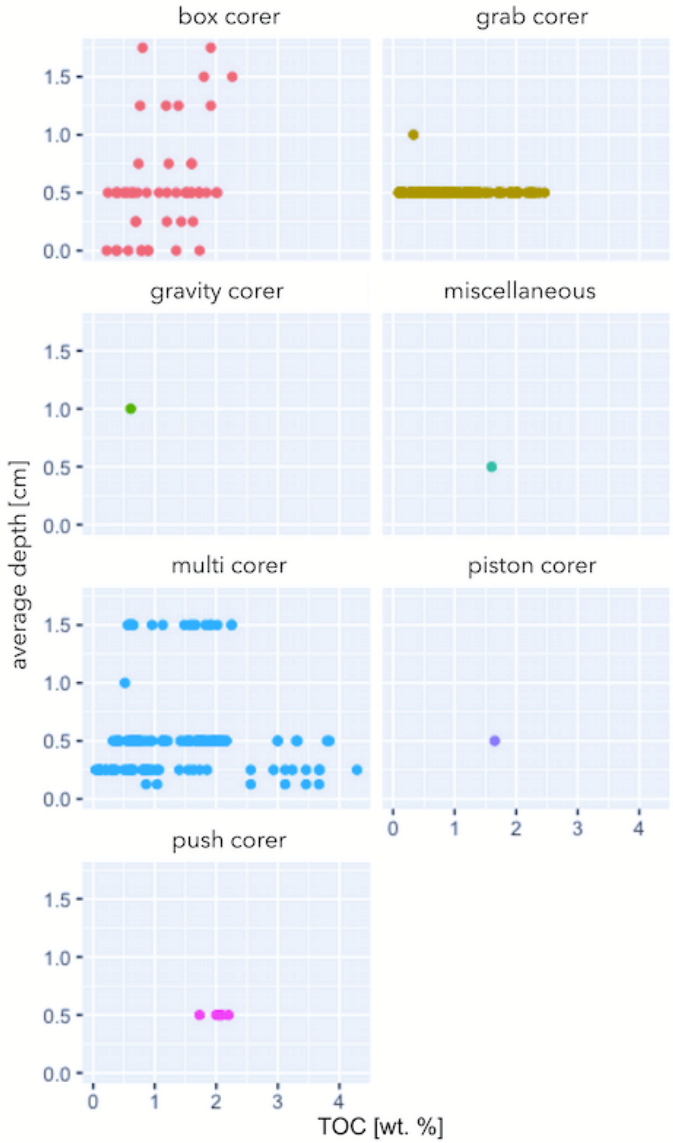

Figure 4.3 TOC content with depth for different sampling methods

In view of the fact that upper and lower sample section depth attributes were sometimes missing, *virtual* upper and lower sample section depths were derived by adding or subtracting, respectively the magnitude of the average sampling interval at a location to/from the local average sample depth.

A custom algorithm derived two datasets (

Table 4.2): one branch preselected all samples having a lower sample section depth <= 2 cm (h2sel2algo2), the other branch preselected samples with an average sample section depth of <= 2 (h2sel1algo2) cm. Where multiple samples at a certain station remained after preselection, the TOC content of that station was calculated based on a weighted mean: The TOC value (respectively average TOC value, if replicates present) of a sample within the top 2 cm was multiplied with its respective sample section magnitude (in cm) which resulted in a weighted mean for each station. The top 2 cm as delineation for surficial was a practical decision, based on a preliminary inspection of the lower sample depths (of all of the samples after adding virtual section limits where needed): most samples where within within a maximum lower sample section depth of 2 cm (Figure 4.4). For a description of the algorithm refer to

Table 4.2 or see code in appendix A.1.

Table 4.2 Description of algorithms for the derived datasets

derived dataset name	Description
h2sel1algo2	Algorithm selects all samples with average sample section depths ≦ 2 cm, and calculated weighted averages for the TOC content.
h2sel2algo2	Algorithm selects all samples with lower sample section depths ≦ 2 cm, and calculates weighted averages for the TOC content after computation of virtual section depths where lower and upper sample section depths were missing.

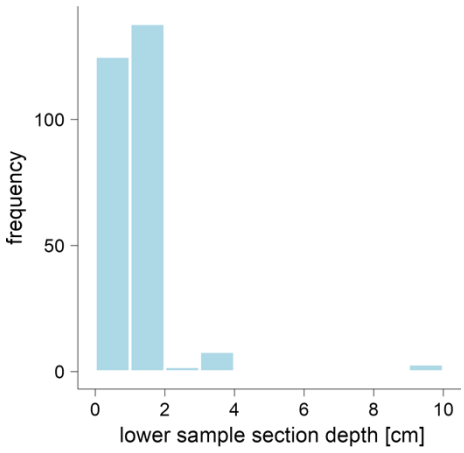

Figure 4.4 Histogram of lower sample section depths after adding virtual sample section limits

4.1.2 Predictor variables

Building upon chapter 3, factors potentially influencing the TOC content in surficial sediments in the study environment can be preselected: bathymetric depth and derivatives (slope, curvature, aspect, roughness of the surface, etc.) distance to coast (also taking into account some measure of deviation based on surface currents), distance to different rivers, distance to the canyon, distance along the canyon axis, primary productivity (and related: chlorophyll a) and speed of bottom currents. From general non-canyon-specific literature, it is known that bottom water temperature (Diesing et al., 2017) and bottom water oxygen concentration (Paropkari et al., 1992) can play into the preservation of sedimentary organic carbon. Only a selection of covariates, already suggesting some control on the TOC content in surficial sediment, will be included in the thesis. The 25 potential covariates were in raster form, for an overview refer to Table 4.3.

4.1.2.a Bathymetric depth and derivatives

The high-resolved bathymetry raster (200m x 200m) made available by the Hydrographic Institute of Portugal is the same one used in the study by Masson et al. (2011). Slope and mean curvature were calculated using the surface parameter functions in the Spatial Analyst Toolbox within ArcGIS Pro. Statistical aspect and rugosity were derived using functions of the Benthic Terrain modeler toolbox (BTM) in ArcMap. Finally, the topographic wetness index (TWI) was computed using the SAGA Module *Wetness Index* within QGIS.

The values of the statistical aspect were transformed using the sine (eastness) and cosine (northness) respectively, to prevent the problem of directions like 359° and 1° being in proximity in the physical world, but remote value-wise. Therefore a trained machine learning model will perform better if the values of north and south, respectively east and west are on divergent scales (-1,+1).

4.1.2.b Distance covariates

Distance to river mouths (see Figure 4.5) and to shoreline was computed with the distance accumulation algorithm that is implemented in ArcGIS Pro. The rivers were extracted as GeoJSON files from OpenStreetMap using the web-based data mining tool Overpass Turbo (overpass-turbo.eu).

The shoreline shapefile for the distance to coast calculation, was downloaded from the World Vector Shorelines (WVS) database which is incorporated in the Global Self-consistent, Hierarchical, High-resolution Geography Database (GSHHG) (Wessel & Smith, 2017).

For the computation of the distance to coast, taking into account the direction of the surface ocean currents, a horizontal factor based on a surface current direction raster (see 4.1.2.c), was defined for the surface accumulation function.

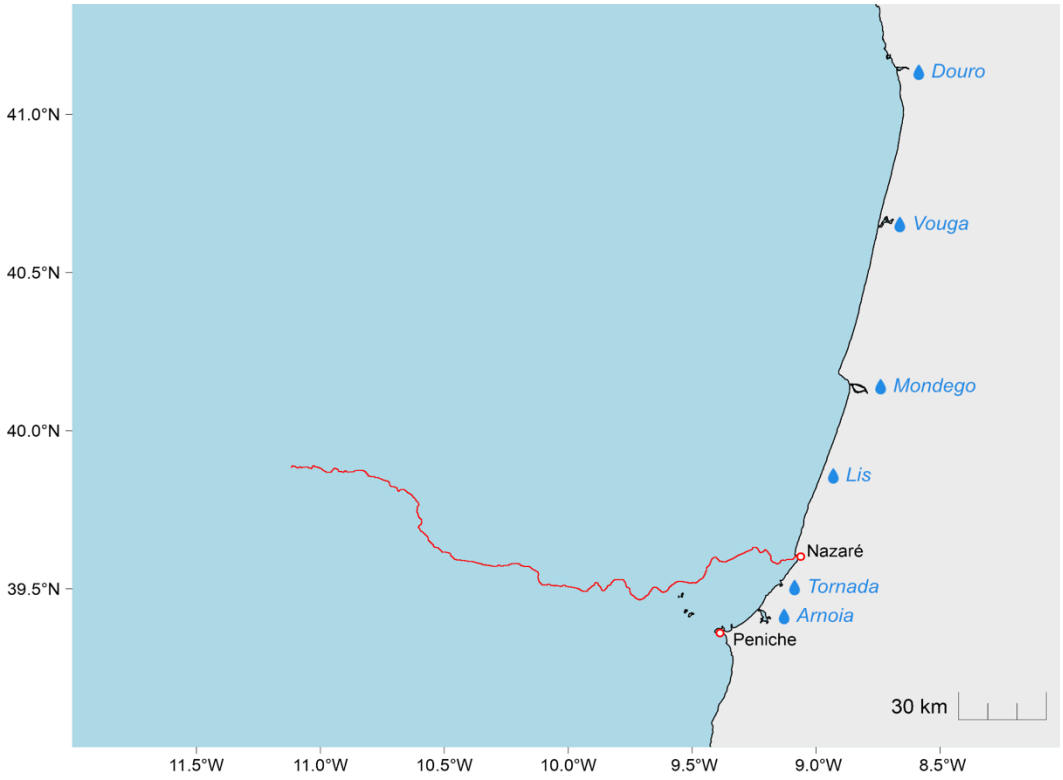

Figure 4.5 Mouth of rivers (blue drops) along the Central Portuguese continental margin

In order to obtain reliable distances, all vector data the distances were calculated to, was reprojected to a custom projected coordinate system, obeying the constraint of equidistance and being optimized for the study region (appendix A.2).

The distances to and along the canyon axis were derived with the *locate features along route* tool within ArcGIS Pro. The canyon axis (or thalweg), acting as the route, was extracted from the bathymetry raster, using an algorithm that was initially developed to extract glacier basins in order to obtain glacier mass balances from digital elevation models. The underlying concept has been described by Bolch et al. (2010) and Kienholz et al. (2013) and has been refined and improved by Philipp Rastner as documented in Falaschi et al. (2017). In order to use this approach with the bathymetry, all depth values were inverted (negative values were made positive, i.e. valleys turned into ridges) and the algorithm derived contiguous polygons which could then be subsequently cleaned manually in a desktop GIS to obtain a single thalweg (Figure 4.6). The use of the generic *flow direction* and the *flow accumulation* functions in ArcGIS did not produce a satisfactory result, i.e. one the thalweg could have been extracted from.

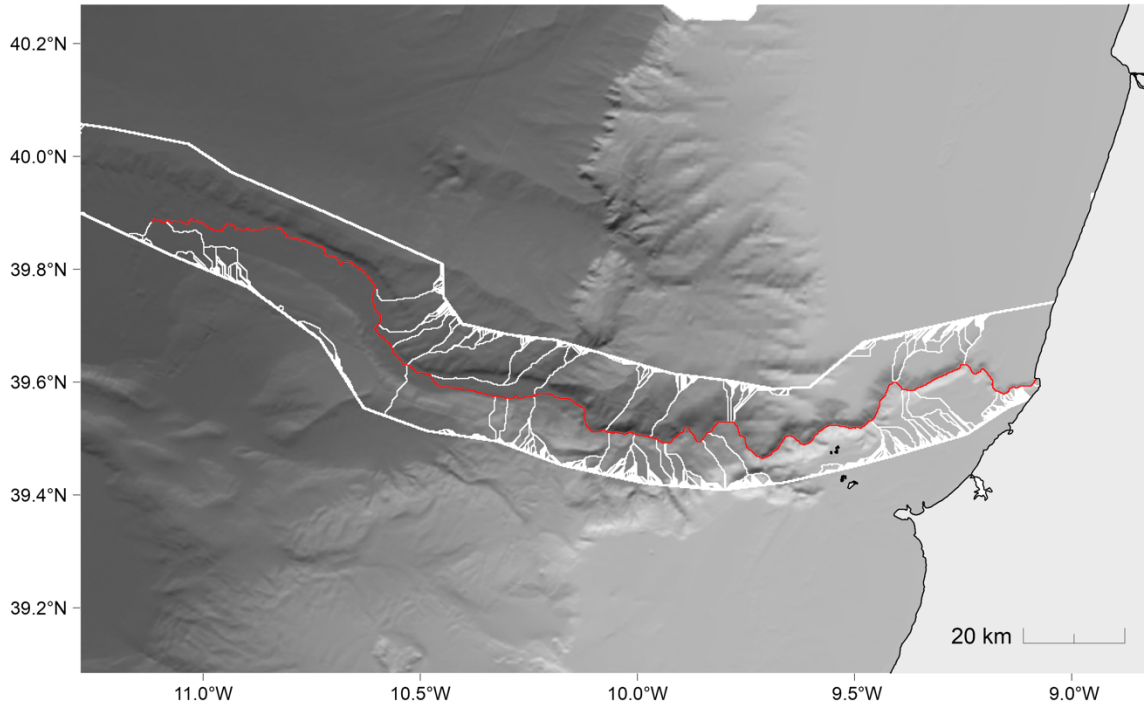

Figure 4.6 Drainage basins (white multipolygon geometry) and the manually extracted thalweg (red polyline)

4.1.2.c Satellite products

Surface ocean (top 0.5 meters) Chlorophyll a and surface ocean net primary productivity as well as bottom water oxygen content were obtained from the Atlantic-Iberian Biscay Irish-Ocean BioGeoChemistry NON ASSIMILATIVE Hindcast for the time period between 15.2.93 to the 15.2.2019. The data consisted of monthly averages, which were then averaged again over the chosen time period to obtain a single mean value. Missing raster cells at the coast were interpolated using the raster calculator and focal statistics, taking the mean over a circular neighbourhood. As in coastal areas, the gradient of the values of interest are much bigger than further apart in the open ocean, the interpolation procedure was repeated several times, taking a radius of only one cell each time to prevent smoothing out the values.

Surface (top 5 meters, Figure 4.7) and bottom currents (Figure 4.8), as well bottom potential temperature were taken from the monthly averaged Global Ocean Physics Reanalysis. Data were averaged over the time period of 1993 to 2016. The net current was calculated on the basis of the eastward and northward current component that were given in this dataset,

Bottom values were always taken as the value at the lowest depth layer, that was not being labelled as NA.

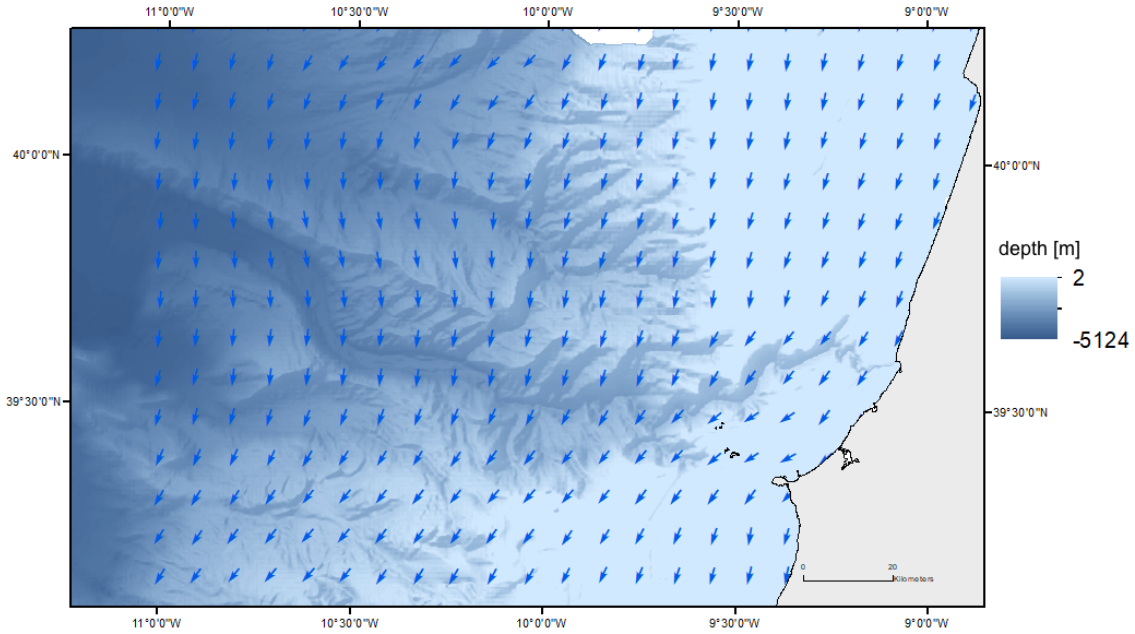

Figure 4.7 Surface current direction (blue arrows)

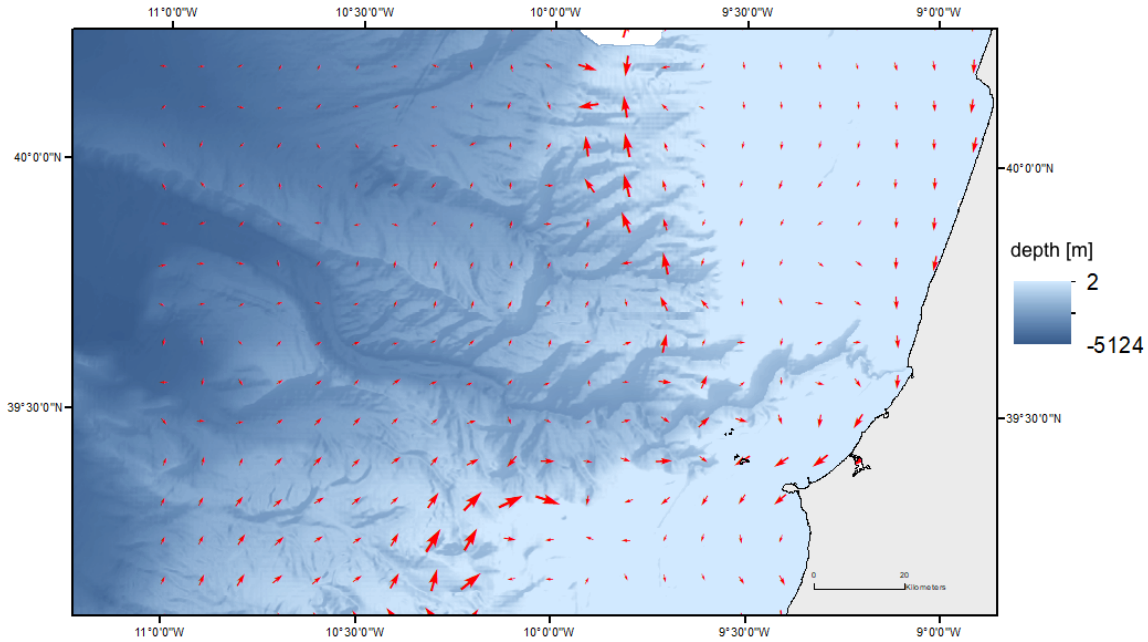

Figure 4.8 Bottom current direction (red arrows) and the relative magnitude of the current (size of arrows)

Table 4.3 Data used as predictor variables in random forest and as coefficients of the linear trend model for kriging with external drift

Covariate	Spatial resolution	Data source	# of covariates
Distance to river mouths: Arnoia, Douro, Lis, Mondego, Tornada, Vouga	20 m	openstreetmap.org	6
Distance to mainland coast – Euclidian – Euclidian normalized by bathymetric depth – cost-based on surface current direction	20 m	Full (f) resolution boundaries between land and ocean (L1) from *World Vector Shorelines (WVS)* database within *Global Self-consistent, Hierarchical, High-resolution Geography Database* (GSHHG), accessed through https://www.ngdc.noaa.gov/mgg/shorelines/ https://resources.marine.copernicus.eu/product-detail/GLOBAL_MULTIYEAR_PHY_001_030/DATA-ACCESS	3
Distance along and to canyon axis	200 m	Canyon thalweg derived with a glacier basin algorithm used on inverted bathymetric depth values (based on an approach by (Falaschi et al., 2017))	2
Bathymetric depth and derivatives – Slope [°] (+ normalized by sine and cosine of statistical aspect) – Sine and cosine of statistical aspect – Mean curvature [°] – Rugosity (VRM=Vector Ruggedness Measure) [0,1] – Topographic wetness index TWI [the higher, the wetter]	200 m	Bathymetry raster from the Hydrographic Institute of Portugal (IHP)	7
– Surface ocean net primary productivity (NPP) in amount of carbon [$mg \cdot m^{-3} \cdot d^{-1}$] – surface ocean chlorophyll a [$mg \cdot m^{-}3$] – bottom oxygen content [mmoles $O_2 \cdot m^{-3}$]	0.083° (~7km at 39° N)	Atlantic-Iberian Biscay Irish-Ocean BioGeoChemistry non assimilative Hindcast (cmems_mod_ibi_bgc_my_0.083deg-3D_P1M-m at https://resources.marine.copernicus.eu/product-detail/IBI_MULTIYEAR_BGC_005_003/INFORMATION)	2
bottom – current speed [$m \cdot s^{-1}$] – potential temperature	0.083°	Global Ocean Physics Reanalysis (GLOBAL_MULTIYEAR_PHY_001_030 at https://resources.marine.copernicus.eu/product-detail/GLOBAL_MULTIYEAR_PHY_001_030/DATA-ACCESS)	3

4.2 Spatial interpolation procedures

Prediction for all the interpolation procedures was performed on a 200 m x 200 m prediction grid (based on the bathymetry raster). The grid nodes correspond to the cell centers of the bathymetry imagery raster cells and were then for mapping purposes transformed to SpatialPixels (a raster equivalent within the spatial "sp" objects in R). For generation of random numbers, the seed was always set to 42.

4.2.1 Kriging with external drift

Geostatistics considers the value of a variable of interest at a certain location (the signal S) as the sum of a realization of a stationary autocorrelated random process Z (with zero mean) and the constant mean μ of the latter. The variance of the random process can be modelled using a variogram.

$$S(x_i) = \mu + Z(x_i)$$

As surficial sedimentary TOC (Figure 5.3 and section 5.1) shows a trend (tendentially higher TOC values within the canyon and lower ones on the shelf) and not a random patchy distribution, we cannot assume that the mean of the spatial random process is constant (which would suggest a stationary process), but changing. Therefore the mean can be portrayed as a deterministic trend component $\mu(x_i)$. As there are covariates d_k at our disposition we can use those to model this trend externally (not based on the response) in a linear regression model. The following model for geostatistical data with a non-stationary mean is assumed (Diggle & Ribeiro, 2007; Nussbaum et al., 2014; Papritz, 2021):

$$Y(x_i) = S(x_i) + \varepsilon_i = \mu(x_i) + Z(x_i) + \varepsilon_i = \sum_k d_k(x_i)\beta_k + Z(x_i) + \varepsilon_i$$

ε is the unresolved iid error, which encompasses for instance measurement errors.

The aim is to have minimum-variance estimates of the trend, unbiased estimates of the residual variogram and finally a known variance for the sum of the trend and the random variation at sites without ground truth. To this end Stein (1999) suggested the use of the empirical best linear unbiased predictor (E-BLUP, a kriging predictor) and a variogram estimated simultaneously with the regression model parameters by residual/restricted maximum likelihood (REML). REML estimates the variance components of the spatial model after having adjusted by ordinary least squares the part of the model containing the fixed (non-stochastic/random) effects (the drift, trend) (Montero et al., 2012).

The residuals of the linear regression model are used to derive the variogram which describes the spatial variation and provides the weights for the kriging interpolator. The predicted value at a given point is then a weighted average of the fitted random function neighboring this point and in the end the trend can be added back to the predictions to get the signal.

First a linear regression model containing all covariates was fitted, and visual analysis of a scatterplot matrix was used to exclude covariates showcasing skewed, narrow or outlier-heavy distributions, indicating they could act as leverage points. Stepwise forward and backward variable selection (based on the Aike Information Criterion AIC or the Bayesian information criterion BIC), as well as exhaustive search selected covariates for the external drift. Analysis of variance (ANOVA, not taking spatial autocorrelation into account) was used to compare candidate models for the external drift and Wald tests (taking spatial autocorrelation into account) and the AIC served to compare the candidate models.

Initial values for the variogram model parameters (e.g., range and nugget) were based on the inspection of the sample variogram of the linear regression residuals and then fitted simultaneously with the linear model for the external drift, using a gaussian (non-robust) REML estimation.

Simultaneous estimation of the drift coefficients and the variogram parameters of the spatial random process was achieved with the R package *georob* (Papritz, 2020). The function *georob* outputs a spatial linear model which contains the coefficients of the drift and the variogram object needed for prediction (*predict.georob)*.The accuracy of the spatial linear model is assessed by 10-fold cross-validation using *cv.georob.*

4.2.2 Forest-based regression

Random Forests (RF) is a supervised learning algorithm that provides ensemble predictions of many individual decision trees (on their own being weak learners) to boost the overall predictive accuracy. Each CART (Classification and Regression Trees, (Breiman et al., 2017)) is made up of nodes, branches and terminal leaves which represent the prediction value. The best set of hyperparameters of a random forest is found during tuning, where for different combinations of the former, different forests are grown and then used to predict the value of interest of a hold-out subset of the data used for training. The hyperparameters achieving best values w.r.t. to a certain accuracy metric (e.g., RMSE) are then used for the training of the final model. When training an individual tree, an optimal covariate (or feature) and a threshold value for splitting are obtained at each node in a way that minimizes the variance. This variance minimization is achieved by bagging (bootstrap aggregating, (Breiman, 1996)) which additionally avoids overfitting to a certain extent: Each tree is grown from a subsample of the full training data, drawn at random and with replacement. The final prediction value is then an average of the predicted values of all decision trees in the forest. The relative importance of the variables within the forest can be accessed and offers the advantage of a possible investigation of predictor-response relationships and which factors might drive a spatial process, e.g., the controls on organic carbon content in marine sediment.

Tuning of the model hyperparameters and unbiased accuracy assessment involved k-fold cross-validation. Aiming to achieve a subdivision of the samples into different folds which well represents the full dataset and to always include some of the less prevalent canyon samples (84 vs 180 outside), the data was split into different folds based on a stratification factor (canyon, outside). The factorial attribute *canyon* was assigned to each data point when taken inside the canyon, *outside* if on the shelf, rise, slope or in the abyss outside of the canyon. As the georeferencing is often inaccurate, this step has not been based on the actual reported coordinates within a source paper but on the description of the sample site if available or using the reported water depth.

4.2.2.a Classic random forest

The classic random forest model was implemented with the *ranger* package (Wright & Ziegler, 2017) in R (2021).

The conventional function *tuneRanger* in the *tuneRanger* package only allows for tuning the hyperparameters mtry, min.node.size and sample.fraction (Table 4.4), therefore the final model was tuned based on 5-fold flat cross-validation using an adjusted code excerpt (https://github.com/AleksandarSekulic/RFSI/blob/d778c2c7b65cb86177e097d565dc54e5db696705/prcp_catalonia/4_prcp_case_study_catalonia_RK.R#L1223, lines 1223-1283) from Aleksandar Sekulić in combination with the function *ranger* of the ranger package. The full dataset was used to train the final model in accordance with the modelling approach promoted by the creators of H2O (*H2O.Ai. (2020) H2o: R Interface for H2O.*, n.d.; *H2O.Ai Docs: Cross-Validation*, 2022).

To obtain an unbiased estimate of the model accuracy, nested (Pejović et al., 2018) 5x5 cross-validation with a stratification (canyon, outside) (de Gruijter et al., 2015) for the outer fold was performed using the full dataset using random grid search (Bhat et al., 2018) based on 60 hyperparameter combinations (Bergstra & Bengio, 2012) drawn from the tuning grid. The nested cross-validation was implemented with the function nestcv.train from the nestedcv R package (Lewis et al., 2022), the only package offering this type of validation for a ranger-type random forest model.

One non-spatial random forest was also created based on a subselection of all the covariates. Misleading variables, which could lead to over-fitting, can be removed with forward feature selection (FFS, *ffs* function in R package *CAST*) priorly to model-building and in view to the chosen validation strategy (Meyer et al., 2018). Meyer and colleagues used spatial cross-validation strategies, e.g. leave-location-out (LLO), which are targeted at spatio-temporal data and would be in our case (no temporal dimension used in data) reduced to leave-one-out (LOO) cross validation, which is computationally demanding and was not implemented in the frame of this project. The approach of preventing overfitting is still interesting and therefore FFS was used with stratified k-fold cross-validation, the validation strategy applied during tuning of the non-spatial forest.

Table 4.4 Hyperparameters for nested and flat cross-validation of the non-spatial random forest models

hyperparameter	definition	values for ranger tuning grid	values for RFSI tuning grid
mtry	number of variables to possibly split at in each node	2 to maximum number of covariates used	2 to maximum number of covariates used
num.trees	number of trees to be trained	500	250
sample.fraction	fraction of observations to sample (1 for sampling with replacement and 0.632 for sampling without replacement)	1	sequential values between 0.65 to 1 in steps of 0.05
min.node.size	minimal node size	2, 4, 6, 8	2, 4, 6, 8
splitrule	criterion the split at each node is based on (variance: minimizes variance in the estimated responses)	variance	variance
n.obs	number of nearest neighbors to consider in rfsi algorithm		2, 3, 4, 5, 6

4.2.2.b Spatial random forest

The spatial random forest was implemented with *random forest spatial interpolation* (RFSI, (Sekulić et al., 2020)) which is part of R package *meteo* (Kilibarda et al., 2014). No approach that had explicitly included geographical context into machine learning (e.g. with distance buffers or geographical coordinates as covariates) had used the values of neighbouring locations as covariates. RFSI uses the values and the distances at n nearest (2D Euclidian distance) neighbours as features in the model and adds with that the basic ingredient of kriging and the majority of the deterministic interpolation methods to machine learning.

Stratified tuning was performed using the function *tune.rfsi* within *meteo*. All other steps and strategies used are identical to the non-spatial model, except the variable selection with FFS, which was only done for the non-spatial version, as rfsi is not a listed model

(http://topepo.github.io/caret/available-models.html) within the *caret* (Kuhn, 2008) package, being used in *CAST::ffs.* The nested cross-validation for an estimate of the unbiased model accuracy was obtained with the *cv.rfsi* function. RFSI calls the *ranger* function, is therefore identical to the classic RF approach except for its additional covariates.

4.3 Model evaluation

The model accuracy is evaluated upon the RMSE from the nested (for the random forest models) respectively flat (for the spatial linear model) cross-validation. Nested cross-validation offers the advantage that the entire dataset can be used for the modelling and no initial test dataset needs to be kept aside for independent validation at the end of the modelling procedure. Meyer et al. (2018) suggest a validation strategy, taking into account the spatiality of the data, whereas Wadoux et al. (2021) reject this and are in favour of random validation strategies. Here a stratification factor (canyon, outside) adds the spatial touch to the validation of the random forest models.

Sekulić et al. (Sekulić et al., 2020) additionally suggest a nested cross-validation structure (Pejović et al., 2018) while Wainer and Cawley (2021) see the nesting as "overzealous" for most practical applications.

The mapping accuracy was derived based on the differences between predicted values and ground truth samples. Due to the limited number of samples throughout the study area, an initial test/train split of the full dataset, with a final model derived with the training dataset and an independent validation using the model to predict the test dataset, was not performed, therefore we rely for the predictive accuracy on the fully nested cross-validation.

4.4 Carbon stock of a submarine canyon

The sedimentary carbon stock for the top 2 cm of the Nazare Canyon has been calculated by summation of the carbon stocks of all bathymetry raster cells (n) within the canyon. Each individual total cell stock (Smeaton et al., 2021) had been calculated by multiplying the planimetric area A of the cell of the bathymetry (cm^2) with the depth of the horizon h (2 cm), the mean dry bulk density (dbd) of the stations (0.61 g cm^{-3}) within the canyon and the predicted TOC content in weight percent at each station. The planimetric area of the canyon, i.e. taking into account the slope of the surface at each raster cell of the bathymetry, was obtained dividing the surface area (200 m by 200 m) by the cosine of the slope in radians. The reference area for the canyon was taken as the polygon from the Seafloor Geomorphic Features Map based on a publication of Harris et. al (2014) and summed up to a planimetric surface area of 3155 km^2 (vs 3079 km^2 surface area).

$$total\ canyon\ organic\ carbon\ stock\ [g] = \sum_{p=1}^{n} \left(\frac{A}{\cos(\theta)}\right)_p \times h \times dbd \times TOC_p$$

To allow for a better comparison of the canyon and its adjacent continental margin, the respective summed up surface areas of the canyon and the continental margin (full study area minus canyon area) were divided by their respective total organic carbon stock to obtain mean carbon stocks.

5 Results

5.1 Spatiotemporal variations of total organic carbon

Figure 5.1 and Table 5 show that the two derived datasets are hardly differing, therefore the choice fell upon the slightly bigger dataset h2sel1algo2.

Table 5 Datasets derived by using different algorithm pre-settings

Dataset	TOC [wt. %]			# samples
	mean	median	range	
h2sel1algo2	1.0034	0.8023	[0.0400, 3.9892]	264
h2sel2algo2	1.0049	0.8045	[0.0400, 3.9892]	263

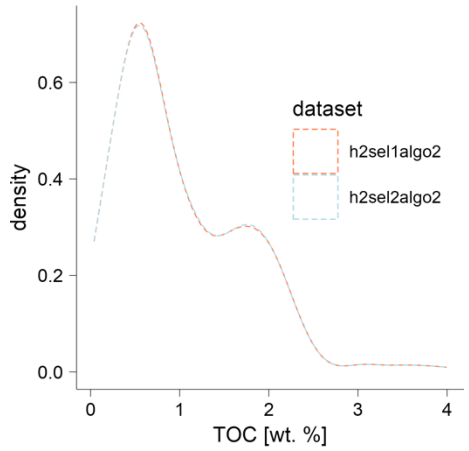

Figure 5.1 Density distributions of TOC content for the two derived datasets

The spatial distribution of the derived weighted mean surficial sediment organic carbon shows a tendency towards higher values within the canyon and its tributaries and lower values outside of the canyon (Figure 5.3). The two settings have their own distinct skewed and bi-modal frequency distributions (Figure 5.2).

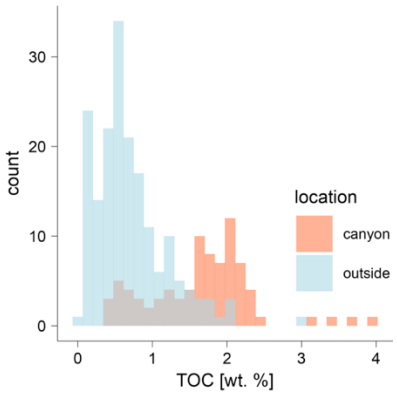

Figure 5.2 Histogram of TOC contents in h2sel1algo2, colored with respect to the location of the data

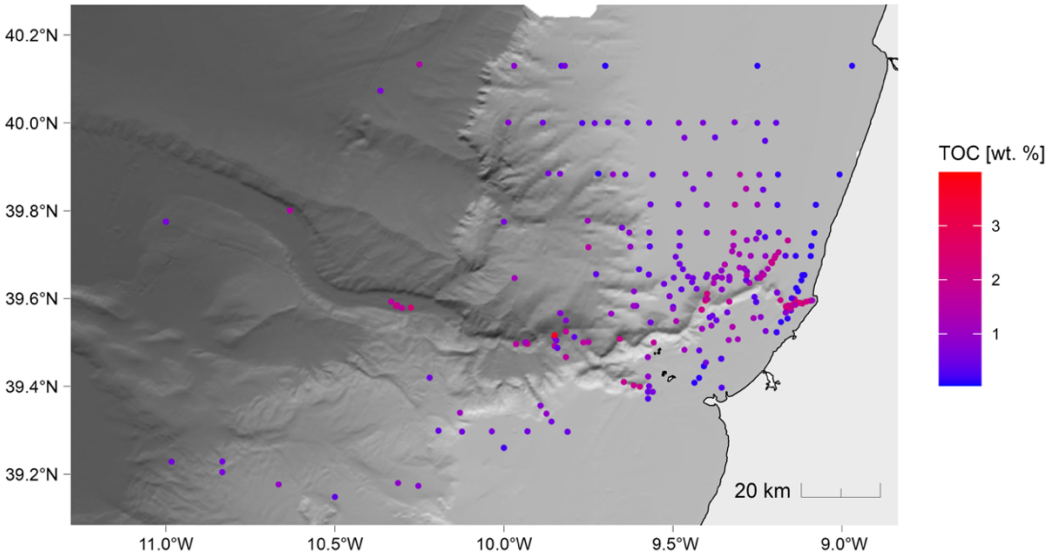

Figure 5.3 Spatial distribution of calculated TOC contents in the dataset h2sel1algo2

Based on reported sediment accumulation rates (SAR) of 0.004-1.453 cm·yr^{-1}, the top 2 cm integrates sediment that has accumulated within years to centuries. The slightly differing ranges of TOC from year to year (Figure 5.4) relfect the spatial varioation of TOC and are not a temporal signal.

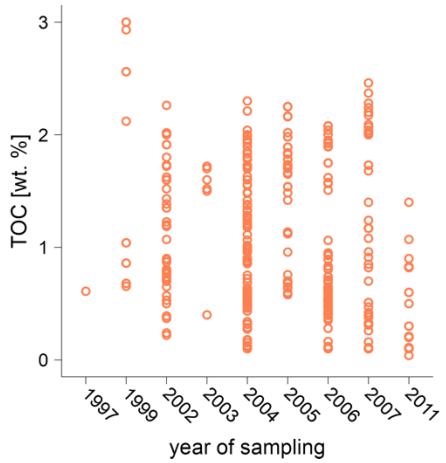

Figure 5.4 Raw (original dataset) TOC content for samples with average section depth of 2 cm

5.2 Modelling and prediction of surficial sediment total organic carbon

Each prediciton approach will be discussed separately, but for better comparison, illustrating figures and accuracy metrics have been arranged side by side.

Table 6 Model performances and predicted TOC content

model	final model for prediction		model accuracy		
	predicted TOC [wt. %]	mapping accuracy	RMSE [wt. %]	MAE [wt. %]	R2
ranger, all features	range: 0.120-2.832 mean: 0.849	rmse: 0.365 mae: 0.229 r2: 0.738	0.534	0.355	0.440
ranger + FFS	0.105-2.876 mean: 0.932	r2: 0.619 rmse: 0.440 mae: 0.307	0.569	0.403	0.363
RFSI	0.135-2.521 mean: 0.895	rmse: 0.369 mae: 0.241 r2: 0.732	0.527	0.358	0.453
external drift kriging	full study area: 1.723×10^{-7} - 21.645 mean full study area: 1.235 canyon: 1.33×10^{-5} - 0.14	rmse: 1.797 mae: 1.141 r2: -5.358	0.705		

5.2.1 Kriging with external drift

As kriging performs best with a normally distributed variable, TOC values were initially transformed using a box-cox transformation of scaled power, which reduced the skewness coefficient from 1.06 to 0.06. As the value of γ (a parameter ensuring that the transformed value is strictly positive) could not be estimated from the data, the log-transform was used instead. This transform still reduced the skewness, with a coefficient of -0.73, but performed worse in doing so than the power transform. Fitting was performed at all steps using gaussian (non-robust) REML.

Rugosity, which was strongly skewed, and slope over the sine and cosine of the statistical aspect, both narrowly distributed and outlier-heavy, were excluded from the modelling of the trend surface (lm1).

An initial stepwise backward selection (based on AIC) on the linear model containing all, but the above removed covariates, resulted in a model (lm2) which did not significantly perform better (p-value from ANOVA 0.65) than the model prior to stepwise selection *lm1.* The full model was kept and then subjected to brute-force (exhaustive) search, meaning that all possible subsets of covariates

were being evaluated. The output contained candidate models from which the best models based on BIC (lm4) and Malow's Cp (lm3) were retained. An omnidirectional sample variogram was then calculated on the residuals from *lm4* using the Mathéron estimator and varying lag distances (bigger lag distances with increasing distance accounted for the sparse sampling in the study area). Subsequently, a first spatial linear model was fitted (sp1), using only the intercept for the drift model (fixed effects) and an exponential variogram model. Stepwise forward feature selection (within the scope of the covariates of lm3) on *sp1*, returned an updated model (sp2), which performed better (based on log-likelihood and AIC) than the intercept-only model (sp1).

The covariates for the external drift in the spatial linear model were then replaced by the ones from *lm4* (resulted in new model sp3) and the new model subjected to stepwise backward feature selection. Pairwise Wald tests found *sp3* to be superior to *sp1* and *sp2*. The coefficients of the drift and variogram parameters of the final model (spx) were therefore based on *sp3* and the exponential variogram model replaced by a Matérn one, which seemed to fit better the sample variogram (Figure 5.5). The smoothness ν of the stochastic process was fixed to 1.5. For an overview of the estimated coefficients of the trend and the parameters of the fitted variogram model refer to Table 7. The model accuracy was assessed by 5-fold cross-validation (Table 6).

The linear kriging interpolator with an external trend surface ranks last in comparison with the other prediction approaches with respect to both the predictive accuracy as well as the mapping accuracy of the model.

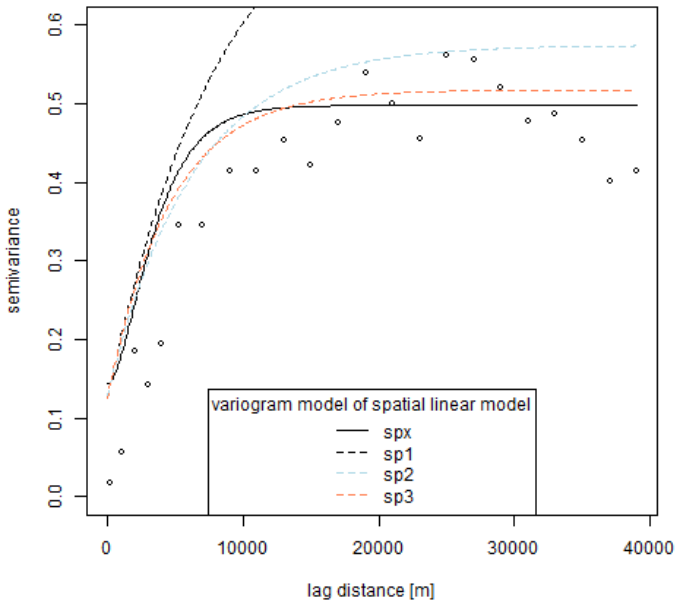

Figure 5.5 Sample variogram (black empty circles) and different fitted variogram models. spx is the final chosen spatial linear model with black, solid variogram model line.

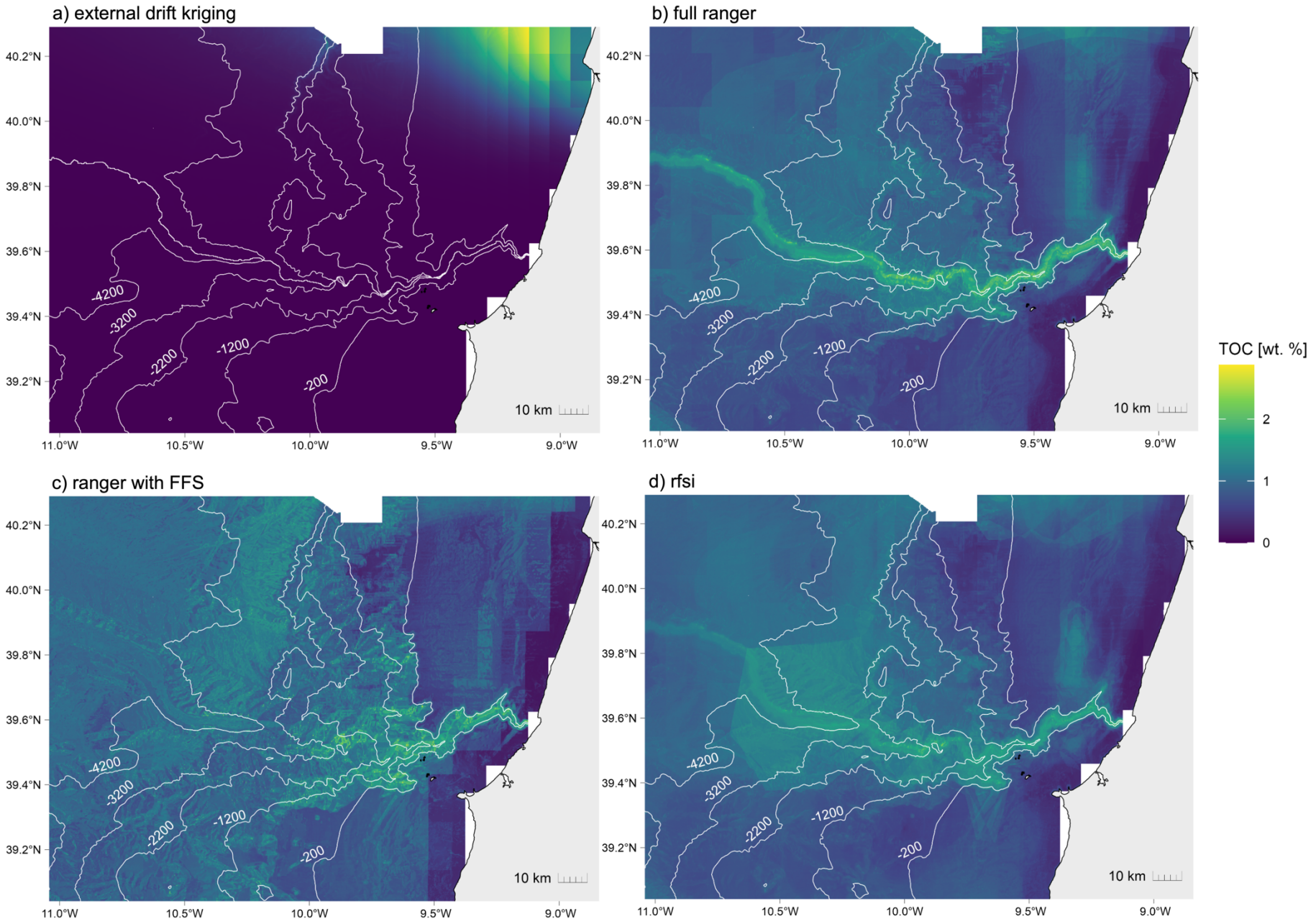

Figure 5.6 Prediction surfaces of TOC. White, solid lines are contours [m]

Table 7 Spatial linear model used for external drift kriging predictions

estimates	value
regression coefficients of drift (fixed effects)	intercept: -0.38 nppv: -0.04 slope: 0.03 IHP: -0.0003 tornada: -0.0001 arnoia: 0.00009 lis: 0.00003 NOAA_mainland: -0.00003
variogram model parameters	variance: 0.35 nugget: 0.14 scale (range of spatial autocorrelation): 3279 m

The range of the predicted values goes largely beyond the one of the available ground truth (0.04 to 3.99 wt. %), with a maximum predicted TOC signal of 21.65 wt. % and a minimum value in the decimals (1.72×10^{-7}). High TOC contents are limited to areas north of the canyon (Figure 5.6a). The upper right corner of the study area exhibits a pattern similar to a multiple ring buffer, with maximum TOC contents in the center and concentrical rings of decreasing values around it. Intermediate TOC values can be found around the northern parts of the foot of the continental slope (FOS).

Zooming into the canyon (clipping kriged prediction surface to Harris' (Harris et al., 2014) canyon outline, Figure 5.7) and constraining the color scale to an upper limit of 0.05 wt. % allows to distinguish more spatial variation by eye. Overall, predicted canyon TOC contents are higher with increasing distance from the canyon head and as well along the canyon flanks.

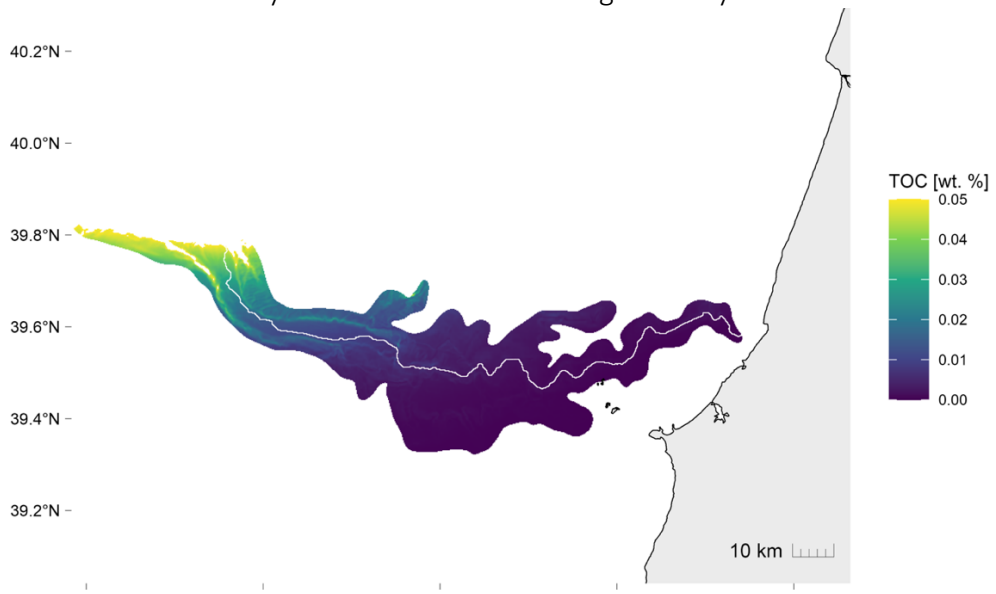

Figure 5.7 External drift kriging predictions clipped to the Harris outline of the Nazaré Canyon. White, solid line is the canyon thalweg (axis).

The model overestimates TOC content mostly where real TOC content is low (< 1wt. %), but underestimates without discrimination low and high TOC contents (Figure 5.8a). Predicted values inside the canyon are always underestimating the observed values. The residuals are spatially

structured with only negative values inside the canyon and its tributaries and positive values at the northern border of the study area (Figure 5.11a). The kriging standard error as an estimate for prediction uncertainty is highest at the coast south to the canyon and then decreases steadily towards the canyon and beyond, before increasing again towards the ring buffer like structure with the highest predicted TOC contents (Figure 5.9).

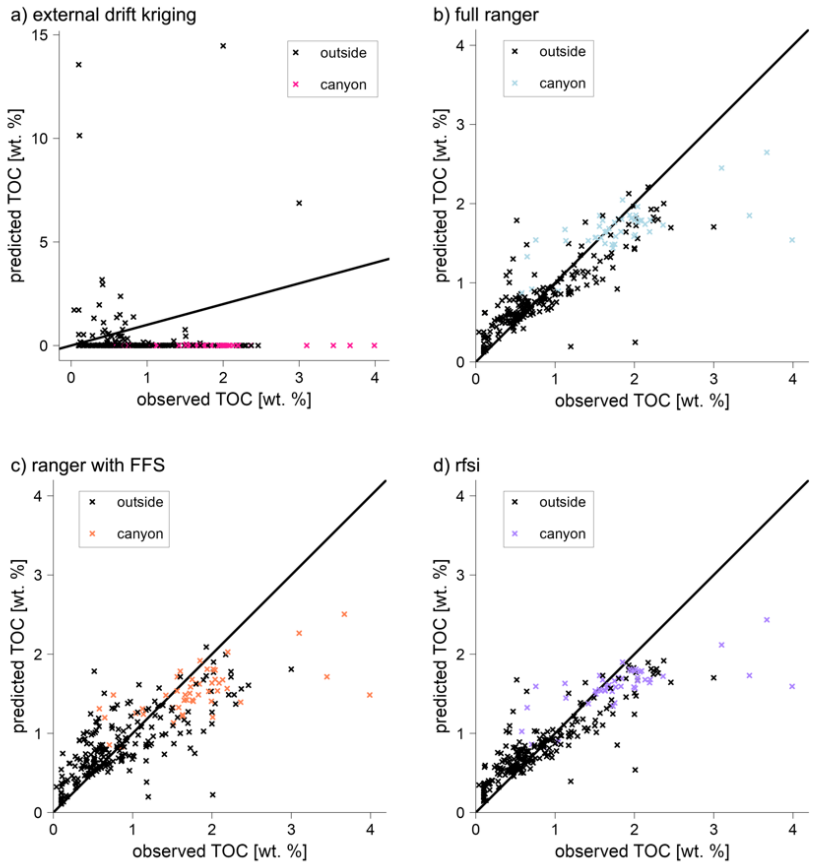

Figure 5.8 Observed versus predicted TOC content (black crosses for locations outside the canyon, colored crosses for locations inside the canyon) with unity line (black, solid line). Observed TOC refers to the derived weighted average TOC contents at each station and not the raw measurement data.

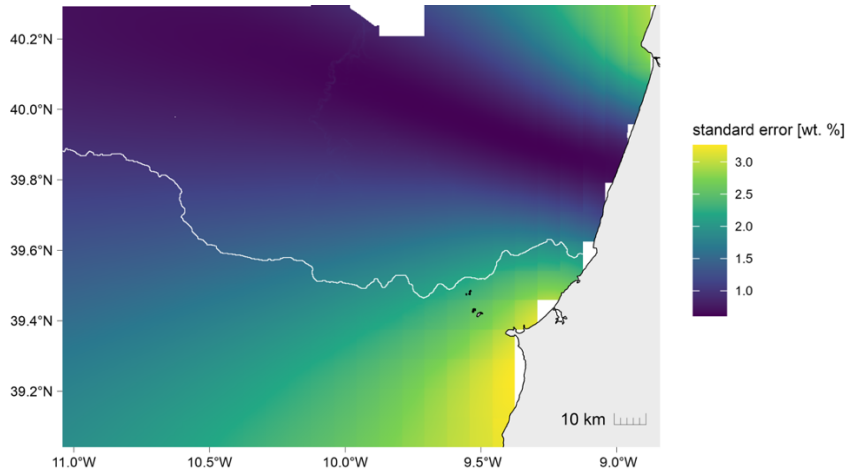

Figure 5.9 Standard error of the external drift kriging predictions with canyon axis (white solid line).

5.2.2 Classic random forest

5.2.2.a All covariates

The full ranger models accuracy comes second amongst all model accuracies, but the model obtained the best mapping accuracy (RMSE= 0.365 wt. %). The model accuracy amounts to an RMSE of 0.534 wt. % almost as good as the RFSI model (RMSE=0.527 wt. %). The most important feature in the forest is the distance to the canyon axis, followed by the distance to the mainland normalized by bathymetric depth and then the rugosity of the terrain. The relative importance (adding up to 100%) of all features in the model are shown in Figure 5.10.

Predicted surficial sediment TOC across the study area (Figure 5.6b) is spatially variable and highlights clearly the canyon. Highest contents are mapped within the latter and its tributaries. Intermediate values are found within a fan-shaped area, spreading from the canyon head and becoming larger towards the abyss. Along the slope, the mapped TOC contents seem to emphasize gully-like structures. Linear artefacts are present on the northern side of the canyon.

Around 44% of the variance in TOC contents could be explained by the model (R^2=0.44). Where the ground truth organic carbon content was low (< 1 wt. %), there was tendentially an overestimation of predicted values, with increasing observed values the underestimation increases as well (Figure 5.8b). Underestimation of high observed values are predominantly in-canyon samples, whereas underestimation of lower observed values, are mostly samples of the adjacent margin. Biggest absolute residuals are found within the canyon and smaller ones on the margin, but there is no clear pattern hinting at a systematic structure (Figure 5.11b).

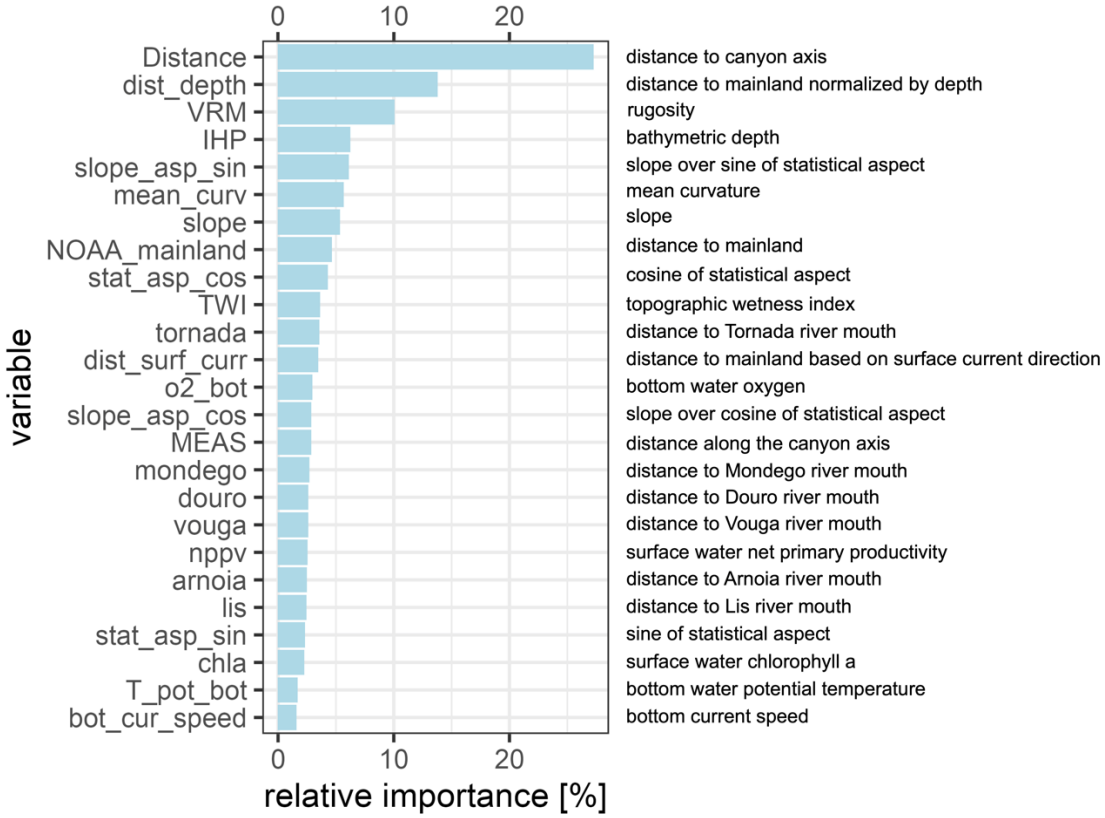

Figure 5.10 Variable importance of the predictor variables indicated by the classic random forest model, using all available covariates.

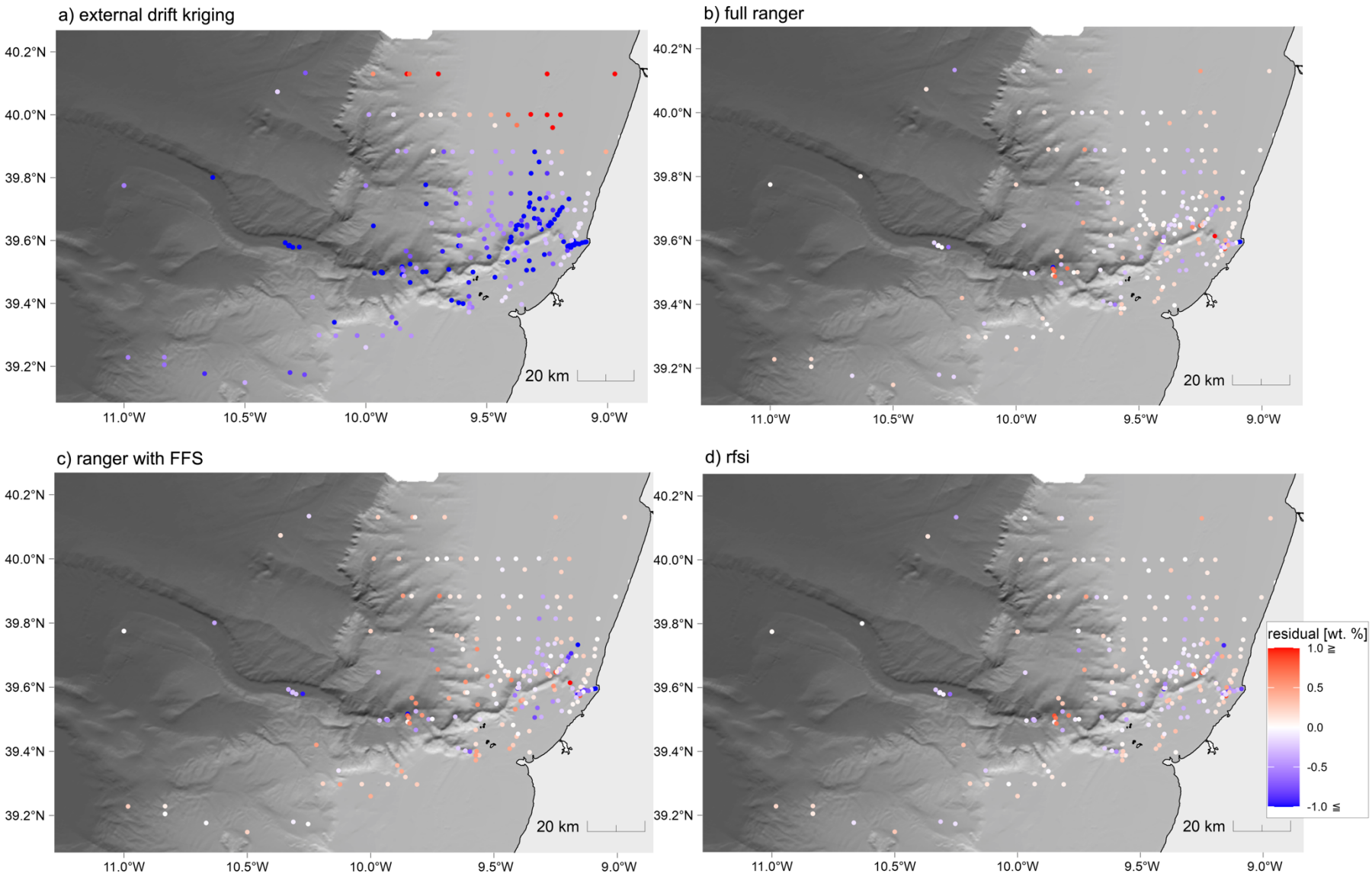

Figure 5.11 Spatial distribution of residuals (predicted value minus observed value)

5.2.2.b Forward selected covariates

Forward feature selection removed 19 of the initial 25 covariates. Rugosity and slope over the sine of the statistical aspect were deemed most important, the remainder of the model features equally dispute their rank (Figure 5.12). The model performs markedly worse (RMSE=0.569 wt.%) than the non-spatial random forest with all covariates. Only 36% (R^2=0.363) of the variation in TOC can be explained by the model. High TOC content is predicted (Figure 5.6c) on the upper part of the canyon flanks within the middle canyon course and the tributary channels. Features along the continental slope (gullies and rills) as well as the shelf break are being traced by the mapped predictions. Likewise, ridges stand out. Low TOC predictions are found mostly along the coast, with lowest values north of the canyon and locally on parts of the shelf break north of the canyon, around 40 degrees latitude. An abrupt change from low, nearshore values to intermediate TOC values offshore results in a visual artefact, a linear feature discontinuously dividing space.

The residual distribution (Figure 5.11c) is similar distribution to the one of the other ranger model, but with some more pronounced negative values around the canyon head, which can also be captured in the scatterplot (Figure 5.8c) where the spread around the 1:1 line is bigger in both canyon and outside points.

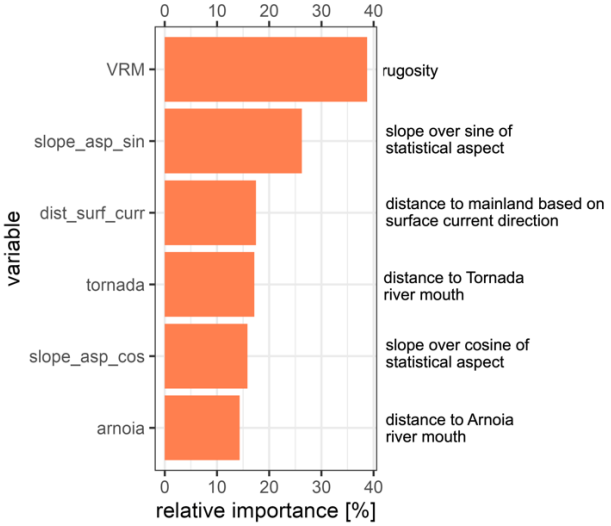

Figure 5.12 Variable importance of the predictor variables indicated by the classic random forest model, using only covariates selected by forward feature selection.

5.2.3 Spatial random forest

The random forest spatial interpolation performs better than the full ranger model, with a slightly smaller RMSE (0.527 wt. %, Table 6). 45% (R^2=0.453) of the variation in measured TOC could be captured by the model. The most important feature is, as in the other random forest models, the distance to the canyon axis, closely followed by the TOC value at the nearest location. Among the most important covariates (Figure 5.13) figure as well distance to the shoreline normalized by depth, rugosity of the terrain and the TOC content of the second closest location.

Predicted TOC values are high within the canyon, as well as in a hexagon-shaped feature reaching from the lower canyon course to the abyss (Figure 5.6d). Intermediate TOC values can be found north of the canyon at the margin of the study area and in an elongated feature on the middle shelf. Lower values are predominantly near the coast and at the northern shelf break. Polygon-like artefacts can also be found on the southern shelf next to a tributary channel of the canyon.

Observed versus predicted TOC values (Figure 5.8d) paint a pattern similar to the one from the non-spatial random forest model with all covariates.

The residuals seem to be distributed randomly with positive and negative ones within the canyon as well as on the adjacent shelf and slope (Figure 5.11d).

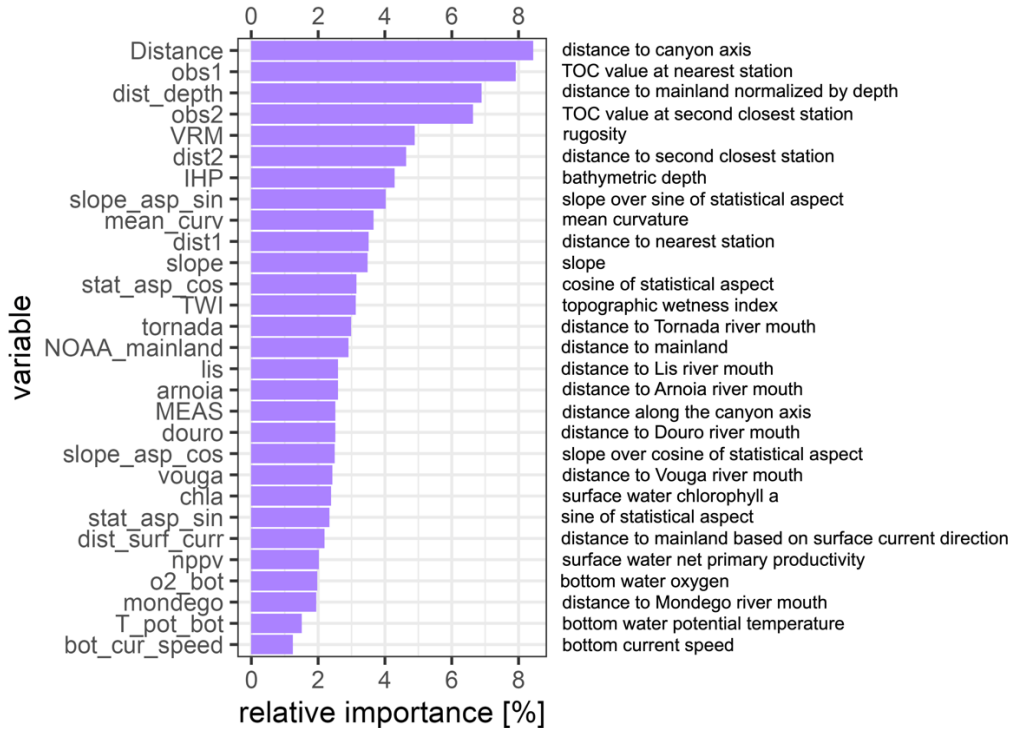

Figure 5.13 Variable importance of the predictor variables indicated by the spatial random forest model (RFSI), using all available covariates.

5.2.4 Areas of enhanced differences

Subtracting prediction surfaces from one another and examining the absolute differences allows for more immediate spatial comparison of the different random forest models. Enhanced differences between the two ranger models (Figure 5.14a) can be found in the middle and lower course of the canyon, as well as along the shelf break and to some extent where the sampling is scarce.

Contrasts between the full ranger model and the RFSI model (Figure 5.14b) are less pronounced than in between the ranger models. Biggest differences are found within the canyon and in the hexagon-like shaped area, an artefact of the RFSI model.

Comparing prediction surfaces of the ranger with preselected covariates and the spatial random forest model (Figure 5.14c), high deviations can be detected along the shelf break, ridges and gullies, as well as within the area delineated by the hexagon artefact. Small differences are found on parts of the northern, well-sampled shelf and the scarcely sampled southern one.

5.2.5 Carbon stock of the Nazaré submarine canyon

The biggest carbon stock for the surficial 2 cm and over the slope corrected study area has been calculated based on the predictions of the full ranger model (0.48 Tg), which is the model with the second-best model accuracy. The most accurate model (RFSI) obtains a slightly smaller stock with 0.47 Tg. The ranger model based on a forward feature selection comes closely after with a stock of 0.46 Tg. For a breakdwon of the stock values refer to Table 8.

Table 8 Top 2 cm total and mean organic carbon stocks

	total carbon stock [Tg = 10^{12} g]			mean carbon stock per unit area [g·m^{-2}]	
model	*canyon*	*outside*	*total*	*canyon*	*outside*
ranger, all features	0.48	1.95	2.43	153	96
ranger + FFS	0.46	2.21	2.67	146	109
RFSI	0.47	2.09	2.56	151	103
external drift kriging	0.02	3.49	3.52	9	170

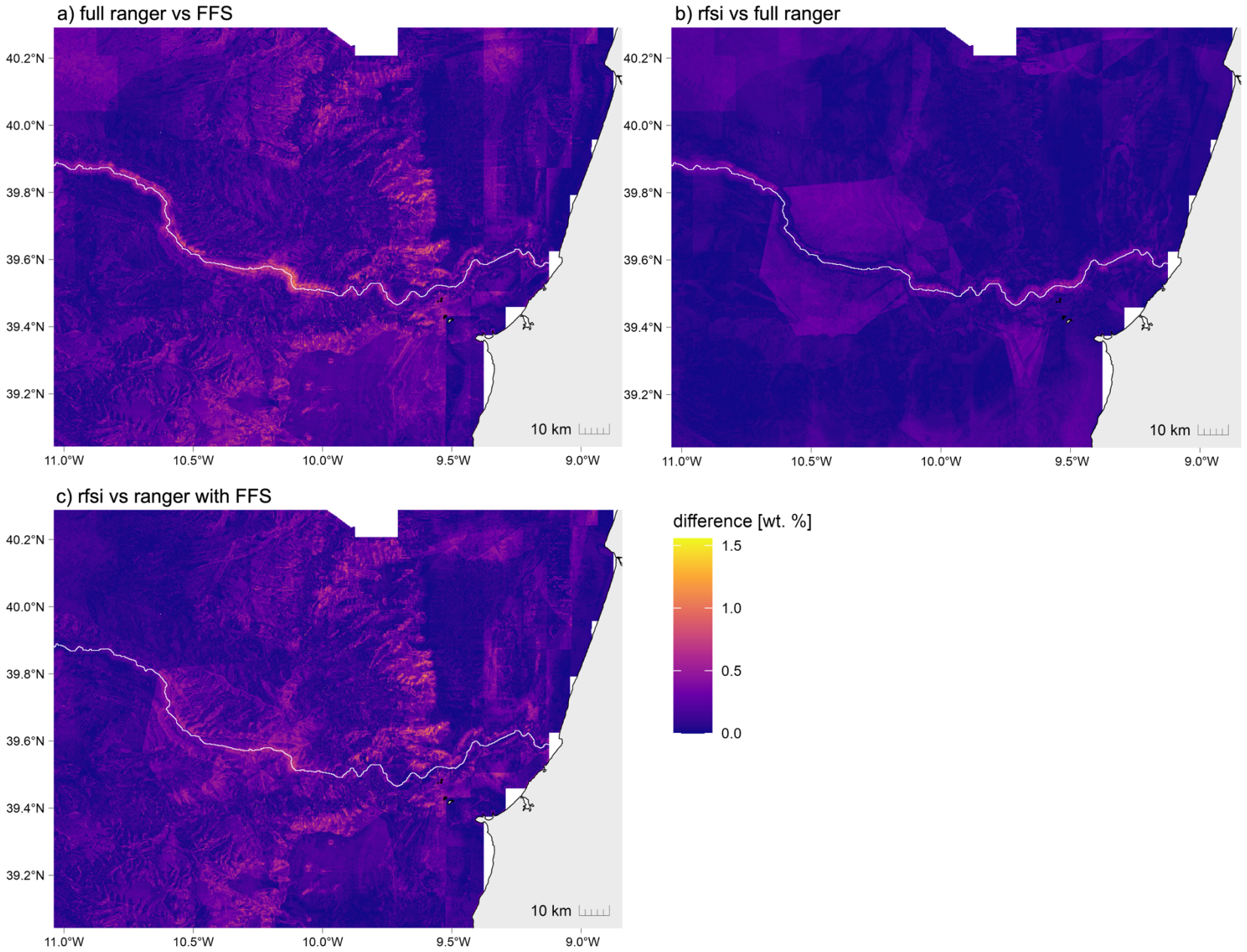

Figure 5.14 Absolute differences between the prediction surfaces of the random forest models. White, solid line is canyon thalweg (axis).

6 Discussion

6.1 Predicting surficial sediment TOC in a heterogeneous setting

Different interpolation methods to predict sedimentary organic carbon within a submarine canyon and the adjacent continental margin have been applied. The results revealed the potential of machine learning for spatial prediction in continental margin settings incised by a submarine canyon. All three random forest models (RMSE between 0.527 and 0.569 wt. %) performed better than kriging with external drift (RMSE=0.705), and the discrepancy might have been even bigger if the spatial linear model had been subjected to nested cross-validation as well to assess its goodness-of-fit (Pejović et al., 2018). Within the random forest models, random forest spatial interpolation (RFSI) performed best, closely followed by the classic ranger model using all the available covariates. The addition of spatial covariates like the distances to closest neighours and the neighbouring TOC values only slightly improved the model accuracy. Visually, all three RF models, show geometrical artefacts, whereby the most prominent one in RFSI (hexagon) seems to stem from the neighbouring points function which will create sharp gradients if the sampling density is low (or distances between two stations too big). The artefacts in the non-spatial RF approaches show a step-like structure this is mostly pronounced in the reduced covariates ranger model along the coast.

The random forest which was trained on a subselection of the available covariates had the lowest predictive accuracy of all three forest models and low explained variance (36%), which could be due to irrelevant highly ranked predictor variables. The forward feature selection was applied once and could have perhaps selected more relevant covariates if applied many times and then using a majority vote to delineate the n most important features (with n being the average number of chosen covariates). Inspecting the density distributions of the predicted TOC contents (Figure 6.1) and the ground truth (serving as training data for the RF and to derive the linear regression model for the drift in kriging), the non-spatial RF based on forward feature selection appears to predict most frequently the mean, acting partially like a null model (a null model (intercept-only model) would systematicaly predict the mean and is thought to perform worse than any model with predictors (Ploton et al., 2020)). This could explain its diminished performance with respect to the other RF models.

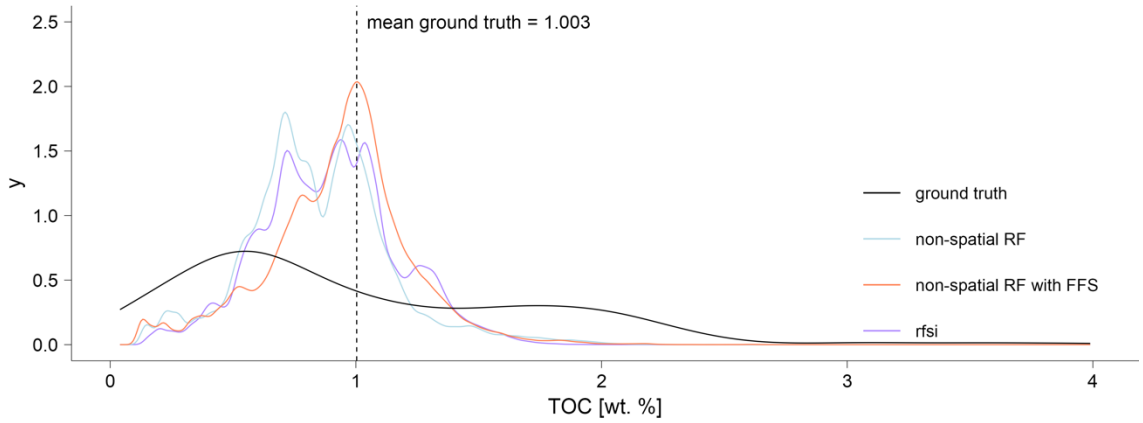

Figure 6.1 TOC content density distributions of full study area prediction surfaces and the ground truth data.

External drift kriging had the lowest predictive accuracy of all approaches and produced a visually very imbalanced representation of predicted TOC. Exploration of the canyon only prediction (Figure 5.7) revealed a pattern suggesting highest TOC contents in the distal canyon. De Stigter et al. (2007) suggested a focused sediment deposition in the middle canyon and flushing events triggered by

turbidity currents and reaching the end of the canyon, occuring on centennial or longer timescales only (not relevant for the scope of this study), there is therefore no physical reason that could partially explain the kriging predictions.

The kriging residuals are the the only ones that show a spatial structure, i.e. no random distribution. This spatial dependence among the residuals and the inherent non-stationarity of the regression coefficients could not be prevented although the variogram model was fitted based on exploration of the residual sample variogram (Finley, 2011). To potentially improve the kriging approach, a directional variogram model could be modelled. The lack of spatial structure in the RF model can be seen as proof that the modelling account for the spatial autocorrelation of the residuals.

Finally, the models accuracy is based on the models ability to predict TOC for the full study area, which was by default set to the greatest common area of all the covariates, being constrained by the locally available bathymetry raster. If a smaller subarea, e.g. the extent (rectangular bounding box of canyon) with a small buffer would have been chosen as study area, the predictive accuracy might have been better as parts of scarcely sampled areas of the shelf and the abyss would have been excluded.

6.2 Potentially important factors influencing the distribution of TOC

In the best performing model (random forest spatial interpolation) distance to the canyon axis, TOC content of the two closest samples, distance to mainland normalized by depth and rugosity identified as key variables in predicting surficial sediment TOC. In the slightly less accurate full ranger model, distance to the canyon axis ranked as within the RFSI first and was then closely followed by the normalized distance to the mainland, rugosity and bathymetric depth. Comparing those predictors to ones being used in global TOC prediction attempts we can see that the distance to the coast plays a role as well in the study by Atwood et al.(2020), but ranks a lot lower than chlorophyll a, which in the RFSI and the full ranger model has very low importance. The important role of distance to shoreline was also shown in a modelling study for the North-West European continental shelf (Diesing et al., 2017) and can be explained by its role as a proxy for input of potentially organic terrestrial material. Rugosity ranks in all three RF models within the first 5, in the FFS ranger even on top. The importance of surface roughness for the presence of cold water corals in a submarine canyon has been shown for the Whittard canyon in the North-East Atlantic (Pearman et al., 2020) and in general (Guinan et al., 2009). Cold water corals also reside in the Nazaré Canyon (Tyler et al., 2009) and offer a potential habitat to a variety of species whose remnants will also provide to the sequestration of organic carbon on the seafloor.

The distance to the mainland normalized by depth ranks higher than the distance and depth covariates alone. As this ratio captures how quickly the bathymetric depth is increasing starting at the shore line (0 m) it can hint at spots where terrestrial carbon can be potentially efficiently transported into depth without loosing too much during lateral transport from the shoreline.

Of all the six rivers considered, Tornada ranks with respect to the other rivers in all RF models highest and has also the biggest coefficient amongst all rivers in the drift regression model used for kriging. The Tornada river mouth lies closest to the canyon head, is thought to be not important though for sediment input into the Nazaré Canyon, as the important biotite fraction in canyon sediments suggests input from more northern rivers like the Douro (Cascalho, 2019). The surface currents (Figure 4.7) also suggest a net southward transport along the coast. But as correlation does not imply causality, the variable importance in RF does not have to be interpreted in a causal manner, as it has been proven that meaningless predictors (for instance photographies of faces) can predict the spatial distribution of an environmental property (Behrens & Viscarra Rossel, 2020).

Inside the RFSI model, the distance to the second closest station ranked higher in the variable importance than the distance to the closest station, which seems counterintuitive considering Toblers first law stating that nearer things are more similar than distant things. An explanation to the models behaviour in this heterogeneous setting could be the fact that the rfsi algorithm uses 2D Euclidian distance to determine n nearest neighbours of a certain point in space. If that said point is now in plan view only 50 meters apart from another point, but is situated on the edge of the canyon and its neighbouring point at the foot of a canyon wall 200 m below, there is little chance that the point laying below would have had an influence on the TOC content of the upper location.

Bottom water temperature and bottom water oxygen content which are known to influence the content of surface sediments, are not labelled as important in any of the models.

6.3 Effect of geomorphological heterogeneity on local OC distribution

Based on the two most accurate models (RFSI and full ranger) some tendencies when it comes to sites of high organic carbon content could be identified. The canyon thalweg and flanks seem to be hotspots of preserved organic carbon, as well as the full length of the shelf break and upper slope. The organc carbon enriched hexagon shape in the RFSI predictions will be ignored as this seems to be an artefact.

Areas of maximum TOC content in the RFSI and full ranger model (Figure 6.2) match areas of highest mapped sedimentation rates from Masson et al. (Figure 6.3). Masson et al. could infere though only a weak correlation between sediment accumulation rates and OC content of surficial sediments. This is surprizing as the preservation of OC «is believed to be strongly influenced by exposure time to oxygenated water, which should be related to sedimentation rate» (Masson (2010) after Hartnett et al. (1998)). If this observation here is purely coincidental would have to be tested with renewed sampling at these predicted hotspot locations and analysis of local sedimentation rates.

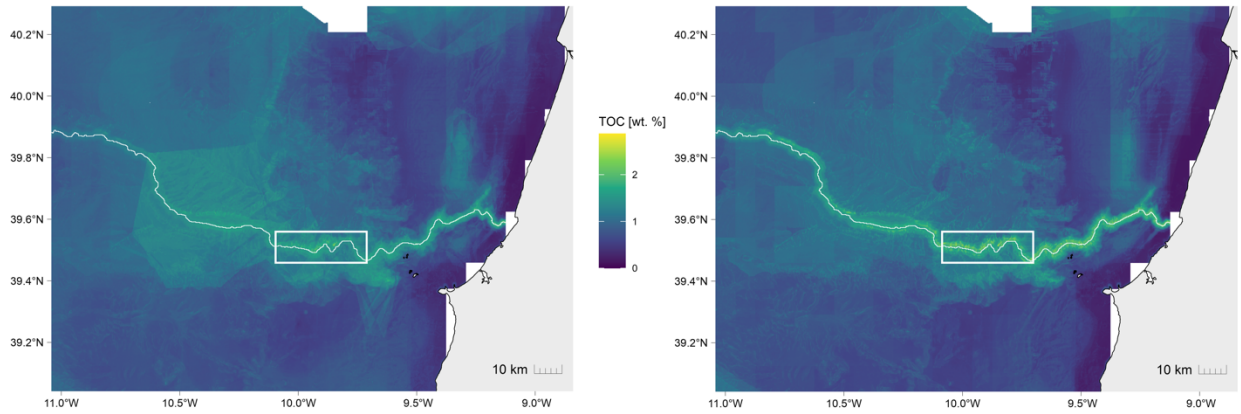

Figure 6.2 Prediction surfces of the RFSI model (left) and the full ranger model (right) with focus box (white outline) and thalweg (white line).

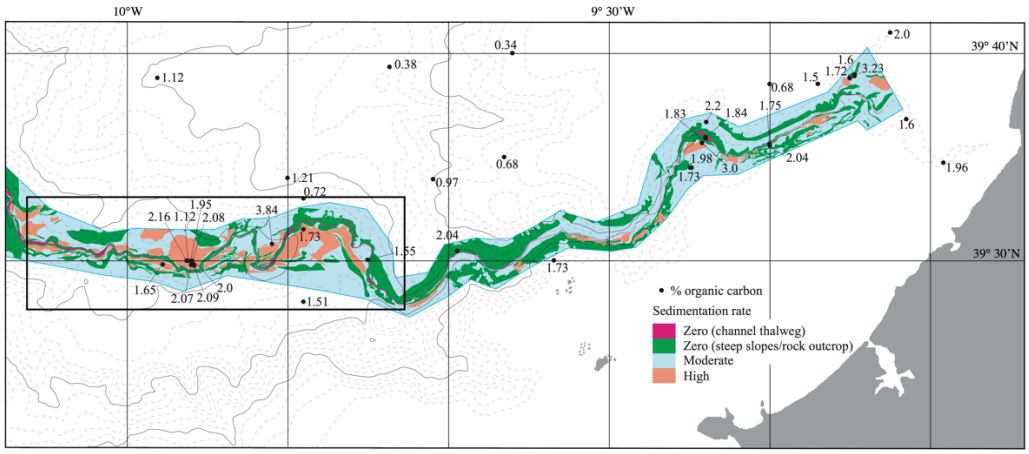

Figure 6.3 OC contents of surface sediments (black points) with color-coded sedimentation rates and thalweg (pink line). Black box added for comparison**.** Adapted from *Efficient burial of carbon in a submarine canyon* by D.G. Masson et al., Geology, 38(9), 831–834.

6.4 Organic carbon stock of a submarine canyon

The highest in-canyon stock has been predicted with the full ranger model, while the prediction surface obtained with external drift kriging has a stock of only 4% of the size of the one from the former. Although the kriging method predicted extremely high TOC values, the total study area stock is only 44% bigger than the smallest total stock (from the full ranger model).

According to the best performing model (RFSI), the Nazaré Canyon can store 0.47 Tg organic carbon in the top 2 cm of its surficial sediment layer. With an organic carbon density of 151 $g \cdot m^{-2}$ it stores 47% more OC in the top 2 cm than the adjacent continental margin. The canyon surface area makes up around 13% of the full study area. If the increased organic carbon storage of canyon sediments is not accounted for, 13% of the study area would be underestimated (by over 50%) when it comes to its surficial organic carbon content. On a global scale, with potentially up to 10% of continental margin area incised by submarine canyons, this would amount to considerable deviations from the real stocks.

6.5 Limitations

A spatial modelling study is dependent on appropriately georeferenced ground truth data. If the positional error is bigger than the spatial resolution of the prediction grid, the predicted value for a certain location might be not representative for the prevailing spatial phenomena. This is even more so a problem in a heterogeneous continental margin setting where small errors can position a sample either inside the canyon, e.g. on a terrace, or on its ridge. Depending on the location the sampled sediment might have been exposed to drastically different environments that could eventually be captured in corresponding covariate values if extracted at the exact position.

The resolution of prediction grid for the RF and the external drift kriging was set to 200 m × 200 m, which corresponds to the resolution of the bathimetry raster. As the covariates used differ in their original spatial resolution (see Table 4.3), they experienced upsampling (were converted to a higher spatial resolution) or downsampling (conversion to a lower spatial resolution) when extracted to the prediction grid. The bias introduced when resampling to a lower resolution (e.g. from the distance accumulation rasters for the river mouths at 20 m × 20 m) is smaller than the other way round. Upsampling an assumption regarding the spatial evolution of a variable at finer scales, which might be not captured by just subdividing a grid cell into equal valued smaller ones. The interpolation of missing coastal raster cells in the raster data of inter alia the net primary productivity raster introduces

bias as well as the mean over a circular neighbourhood might not reproduce a potentially strong gradient in the phenomenon. Aggregation did not only happen on a spatial scale but as well on a temporal one when averaging monthly mean values over the available time range of the processed satellite imagery data for certain covariates.

Another issue is the change of support (Gelfand, 2001). The ground truth data is on point support, but the predicted values are then displayed as a raster (SpatialPixelsDataFrame), i.e. on areal support (predictions have still been made at point support but subsequently been extended to raster cells in order to create a gapless surface). The prediciton grid nodes where then considered as the cell centers of the resulting raster cells.

For the calculation of the organic carbon stock a mean dry bulk density (0.61 g cm^{-3}) derived from a 5 sample locations (where this attribute was given) was used for the full study area. As this quantity is varying within the study area from 0.58 to 0.66, it can be assumed that the effect of this mean value over the whole area can be neglected.

7 Conclusions and Outlook

In this study, different approaches to predict surficial sediment TOC were put to test in a highly heterogeneous setting. Random forest models surpassed the predictive accuracy of the geostatistical model (RMSE= 0.705 wt.%) by far with a RMSE of 0.527 to 0.569 wt.% and up to 45% explained variance. RF models therefore constitute a valid method to predict TOC contents in surface sediments of continental margin systems with incising submarine canyons. To determine the validity and robustness of the approach, additional sampling campaigns would have to be undertaken to compare the predicted values to the ground truth in areas where no sampling had been conducted previously.

To better constrain the spatial limits of predictability, the concept of the area of applicability (AOA) by Meyer et. al. (2021) could be applied to infere how far and where the random forest models are able to predict, based on a given set of spatial covariates. Additionally a more informed use of the forward feature selection (proposed by Meyer et al. (2018)) might effectively reduce the number of irrelevant covariates. For this means, random forest spatial interpolation would need to be added to allowed models within the *caret* package.

The algorithm of the best performing model (RFSI) could be extended, taking into account the configuration of the sampled station and accounting for abrupt landscape evolution, by calculating 3D Euclidian distance as well as query a 3D distance matrix with the respective water depths of the sampled stations. This way only stations would be included as neighbours, if they were suspected to be influential for a location in question, meaning being closer and higher up. But, as there is potentially also up-canyon flow (de Stigter et al., 2011) this would have to be considered as well, by adding a preferential current direction through all depths based on modelled ocean currents taking into account bathymetry (e.g. the Regional Ocean Modeling Sytem (ROMS)). To improve the kriging model and obtain a real measure of how much better a forest model could predict, a directional variogram could be modelled to take into account any anisotropy in the spatial variation.

Finally, to extend the findings of this study, these non-spatial and spatial (rfsi) machine learning approaches need to be applied to different margin canyon systems in order to see how the models using the same covariates would perform in another location.

The present study showed the potential of machine learning in spatial prediction of surficial sediment TOC in a submarine canyon and its adjacent continental margin and potential factors influencing the distribution of the latter. We indentified preferential areas of higher TOC contents and identified the potential OC storage character of Nazaré Canyon (up to 0.48 Tg). Submarine canyons need therefore be accounted for in carbon stocks of continental margin systems.

References

Acharya, S. S., & Panigrahi, M. K. (2016). Evaluation of factors controlling the distribution of organic matter and phosphorus in the Eastern Arabian Shelf: A geostatistical reappraisal. *Continental Shelf Research*. https://doi.org/10.1016/j.csr.2016.08.001

Allin, J. R., Hunt, J. E., Talling, P. J., Clare, M. A., Pope, E., & Masson, D. G. (2016). Different frequencies and triggers of canyon filling and flushing events in Nazaré Canyon, offshore Portugal. *Marine Geology*, *371*, 89–105. https://doi.org/10.1016/j.margeo.2015.11.005

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician*. https://doi.org/10.1080/00031305.1992.10475879

Appah, J. K. M., Lim, A., Harris, K., O'Riordan, R., O'Reilly, L., & Wheeler, A. J. (2020). Are Non-reef Habitats as Important to Benthic Diversity and Composition as Coral Reef and Rubble Habitats in Submarine Canyons? Analysis of Controls on Benthic Megafauna Distribution in the Porcupine Bank Canyon, NE Atlantic. *Frontiers in Marine Science*, *7*. https://doi.org/10.3389/fmars.2020.571820

Arzola, R. G., Wynn, R. B., Lastras, G., Masson, D. G., & Weaver, P. P. E. (2008). Sedimentary features and processes in the Nazaré and Setúbal submarine canyons, west Iberian margin. *Marine Geology*, *250*(1–2), 64–88. https://doi.org/10.1016/j.margeo.2007.12.006

Atwood, T. B., Witt, A., Mayorga, J., Hammill, E., & Sala, E. (2020). Global Patterns in Marine Sediment Carbon Stocks. *Frontiers in Marine Science*, *7*(March), 1–9. https://doi.org/10.3389/fmars.2020.00165

Ausín, B., Bruni, E., Haghipour, N., Welte, C., Bernasconi, S. M., & Eglinton, T. I. (2021). Controls on the abundance, provenance and age of organic carbon buried in continental margin sediments. *Earth and Planetary Science Letters*, *558*, 116759. https://doi.org/10.1016/j.epsl.2021.116759

Avelar, S., van der Voort, T. S., & Eglinton, T. I. (2017). Relevance of carbon stocks of marine sediments for national greenhouse gas inventories of maritime nations. *Carbon Balance and Management 2017 12:1*, *12*(1), 1–10. https://doi.org/10.1186/S13021-017-0077-X

Baudin, F., Martinez, P., Dennielou, B., Charlier, K., Marsset, T., Droz, L., & Rabouille, C. (2017). Organic carbon accumulation in modern sediments of the Angola basin influenced by the Congo deep-sea fan. *Deep-Sea Research Part II: Topical Studies in Oceanography*. https://doi.org/10.1016/j.dsr2.2017.01.009

Behrens, T., & Viscarra Rossel, R. A. (2020). On the interpretability of predictors in spatial data science: the information horizon. *Scientific Reports*. https://doi.org/10.1038/s41598-020-73773-y

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*.

Bhat, P. C., Prosper, H. B., Sekmen, S., & Stewart, C. (2018). Optimizing event selection with the random grid search. *Computer Physics Communications*. https://doi.org/10.1016/j.cpc.2018.02.018

Bianchi, T. S., Cui, X., Blair, N. E., Burdige, D. J., Eglinton, T. I., & Galy, V. (2018). Centers of organic carbon burial and oxidation at the land-ocean interface. *Organic Geochemistry*, *115*, 138–155. https://doi.org/10.1016/j.orggeochem.2017.09.008

Bolch, T., Yao, T., Kang, S., Buchroithner, M. F., Scherer, D., Maussion, F., Huintjes, E., & Schneider, C. (2010). A glacier inventory for the western Nyainqentanglha range and the Nam Co Basin, Tibet, and glacier changes 1976-2009. *Cryosphere*. https://doi.org/10.5194/tc-4-419-2010

Breiman, L. (1996). Bagging predictors. *Machine Learning*. https://doi.org/10.1007/bf00058655

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees [Book]. In L. Breiman (Ed.), *Classification and Regression Trees*. Wadsworth International Group. https://doi.org/10.1201/9781315139470

Cascalho, J. (2019). Provenance of Heavy Minerals: A Case Study from the WNW Portuguese Continental Margin. *Minerals*. https://doi.org/10.3390/min9060355

Chen, L., Ren, C., Li, L., Wang, Y., Zhang, B., Wang, Z., & Li, L. (2019). A comparative assessment of geostatistical, machine learning, and hybrid approaches for mapping topsoil organic carbon content. *ISPRS International Journal of Geo-Information*, *8*(4), 1–18. https://doi.org/10.3390/ijgi8040174

Cordes, E. E., Jones, D. O. B., Schlacher, T. A., Amon, D. J., Bernardino, A. F., Brooke, S., Carney, R., DeLeo, D. M., Dunlop, K. M., Escobar-Briones, E. G., Gates, A. R., Génio, L., Gobin, J., Henry, L. A., Herrera, S., Hoyt, S., Joye, M., Kark, S., Mestre, N. C., … Witte, U. (2016). Environmental impacts of the deep-water oil and gas industry: A review to guide management strategies. In *Frontiers in Environmental Science*. https://doi.org/10.3389/fenvs.2016.00058

de Gruijter, J. J., Minasny, B., & Mcbratney, A. B. (2015). Optimizing stratification and allocation for design-based estimation of spatial means using predictions with error. *Journal of Survey Statistics and Methodology*. https://doi.org/10.1093/jssam/smu024

de Leo, F. C., Vetter, E. W., Smith, C. R., Rowden, A. A., & McGranaghan, M. (2014). Spatial scale-dependent habitat heterogeneity influences submarine canyon macrofaunal abundance and diversity off the Main and Northwest Hawaiian Islands. *Deep-Sea Research Part II: Topical Studies in Oceanography*. https://doi.org/10.1016/j.dsr2.2013.06.015

de Stigter, H. C., Boer, W., de Jesus Mendes, P. A., Jesus, C. C., Thomsen, L., van den Bergh, G. D., & van Weering, T. C. E. (2007). Recent sediment transport and deposition in the Nazaré Canyon, Portuguese continental margin. *Marine Geology*, *246*(2–4), 144–164. https://doi.org/10.1016/j.margeo.2007.04.011

de Stigter, H. C., Jesus, C. C., Boer, W., Richter, T. O., Costa, A., & van Weering, T. C. E. (2011). Recent sediment transport and deposition in the Lisbon-Setúbal and Cascais submarine canyons, Portuguese continental margin. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *58*(23–24), 2321–2344. https://doi.org/10.1016/j.dsr2.2011.04.001

Delaunay, B. (1934). Sur la sphere vide. *Bulletin de l'Académie Des Sciences de l'URSS*.

Delhomme, J.-P. (1979). RetIexions sur la prise en compte simultanee des don- nees de forages et des donnees sismiques. In *Publication LHM/RC/79/41*.

Diesing, M., Kröger, S., Parker, R., Jenkins, C., Mason, C., & Weston, K. (2017). Predicting the standing stock of organic carbon in surface sediments of the North–West European continental shelf. *Biogeochemistry*, *135*(1), 183–200. https://doi.org/10.1007/s10533-017-0310-4

Diesing, M., Thorsnes, T., & Rún Bjarnadóttir, L. (2021). Organic carbon densities and accumulation rates in surface sediments of the North Sea and Skagerrak. *Biogeosciences*. https://doi.org/10.5194/bg-18-2139-2021

Diggle, P. J., & Ribeiro, P. J. (2007). Model-based Geostatistics (Springer Series in Statistics). In *Journal of the Royal …*.

Emerson, S., & Hedges, J. I. (1988). Processes controlling the organic carbon content of open ocean sediments. *Paleoceanography*. https://doi.org/10.1029/PA003i005p00621

Erdogan Erten, G., Yavuz, M., & Deutsch, C. v. (2022). Combination of Machine Learning and Kriging for Spatial Estimation of Geological Attributes. *Natural Resources Research*. https://doi.org/10.1007/s11053-021-10003-w

Escobar-Briones, E., & García-Villalobos, F. (2009). Distribution of total organic carbon and total nitrogen in deep-sea sediments from the southwestern Gulf of Mexico. *Boletin de La Sociedad Geologica Mexicana*, *61*, 73–86. https://doi.org/10.18268/BSGM2009v61n1a7

Falaschi, D., Bolch, T., Rastner, P., Lenzano, M. G., Lenzano, L., lo Vecchio, A., & Moragues, S. (2017). Mass changes of alpine glaciers at the eastern margin of the Northern and Southern Patagonian Icefields between 2000 and 2012. *Journal of Glaciology*, *63*(238), 258–272. https://doi.org/10.1017/jog.2016.136

Finley, A. O. (2011). Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution*. https://doi.org/10.1111/j.2041-210X.2010.00060.x

Fouedjio, F., & Klump, J. (2019). Exploring prediction uncertainty of spatial data in geostatistical and machine learning approaches. *Environmental Earth Sciences*. https://doi.org/10.1007/s12665-018-8032-z

García, R., van Oevelen, D., Soetaert, K., Thomsen, L., de Stigter, H. C., & Epping, E. (2008). Deposition rates, mixing intensity and organic content in two contrasting submarine canyons. *Progress in Oceanography*, *76*(2), 192–215. https://doi.org/10.1016/j.pocean.2008.01.001

Gelfand, A. E. (2001). On the change of support problem for spatio-temporal data. *Biostatistics*. https://doi.org/10.1093/biostatistics/2.1.31

Guinan, J., Brown, C., Dolan, M. F. J., & Grehan, A. J. (2009). Ecological niche modelling of the distribution of cold-water coral habitat using underwater remote sensing data. *Ecological Informatics*. https://doi.org/10.1016/j.ecoinf.2009.01.004

H2O.ai. (2020) h2o: R Interface for H2O. (R package version 3.30.0.6.). (n.d.). https://github.com/h2oai/h2o-3.

H2O.ai docs: Cross-Validation. (2022). https://docs.h2o.ai/h2o/latest-stable/h2o-docs/cross-validation.html

Harris, P. T., Macmillan-Lawler, M., Rupp, J., & Baker, E. K. (2014). Geomorphology of the oceans. *Marine Geology*, *352*, 4–24. https://doi.org/10.1016/j.margeo.2014.01.011

Harris, P. T., & Whiteway, T. (2011). Global distribution of large submarine canyons: Geomorphic differences between active and passive continental margins. *Marine Geology*. https://doi.org/10.1016/j.margeo.2011.05.008

Hartnett, H. E., Keil, R. G., Hedges, J. I., & Devol, A. H. (1998). Influence of oxygen exposure time on organic carbon preservation in continental margin sediments. *Nature*. https://doi.org/10.1038/35351

Hedges, J. I., & Keil, R. G. (1995). Sedimentary organic matter preservation: an assessment and speculative synthesis. In *Marine Chemistry*. https://doi.org/10.1016/0304-4203(95)00008-F

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *2018*(8). https://doi.org/10.7717/peerj.5518

Huvenne, V. A. I., Pattenden, A. D. C., Masson, D. G., & Tyler, P. A. (2012). Habitat Heterogeneity in the Nazaré Deep-Sea Canyon Offshore Portugal. In *Seafloor Geomorphology as Benthic Habitat* (pp. 691–701). Elsevier Inc. https://doi.org/10.1016/B978-0-12-385140-6.00050-5

IPCC Working Group 1, I., Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., & Midgley, P. M. (2013). IPCC, 2013: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. *IPCC*.

Kandasamy, S., & Nath, B. N. (2016). Perspectives on the terrestrial organic matter transport and burial along the land-deep sea continuum: Caveats in our understanding of biogeochemical processes and future needs. In *Frontiers in Marine Science*. https://doi.org/10.3389/fmars.2016.00259

Kharbush, J. J., Close, H. G., van Mooy, B. A. S., Arnosti, C., Smittenberg, R. H., le Moigne, F. A. C., Mollenhauer, G., Scholz-Böttcher, B., Obreht, I., Koch, B. P., Becker, K. W., Iversen, M. H., & Mohr, W. (2020). Particulate Organic Carbon Deconstructed: Molecular and Chemical Composition of Particulate Organic Carbon in the Ocean. In *Frontiers in Marine Science*. https://doi.org/10.3389/fmars.2020.00518

Kienholz, C., Hock, R., & Arendt, A. A. (2013). Instruments and Methods A new semi-automatic approach for dividing glacier complexes into individual glaciers. *Journal of Glaciology*. https://doi.org/10.3189/2013JoG12J138

Kilibarda, M., Hengl, T., Heuvelink, G. B. M., Gräler, B., Pebesmatadić, E. P., & Bajat, B. (2014). Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. *Journal of Geophysical Research*. https://doi.org/10.1002/2013JD020803

Kiriakoulakis, K., Blackbird, S., Ingels, J., Vanreusel, A., & Wolff, G. A. (2011). Organic geochemistry of submarine canyons: The Portuguese Margin. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *58*(23–24), 2477–2488. https://doi.org/10.1016/j.dsr2.2011.04.010

Kopczewska, K. (2022). Spatial machine learning: new opportunities for regional science. *Annals of Regional Science*. https://doi.org/10.1007/s00168-021-01101-x

Krumbein, W. C. (1959). Trend surface analysis of contour-type maps with irregular control-point spacing. *Journal of Geophysical Research*. https://doi.org/10.1029/jz064i007p00823

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5 SE-Articles), 1–26. https://doi.org/10.18637/jss.v028.i05

LaRowe, D. E., Arndt, S., Bradley, J. A., Estes, E. R., Hoarfrost, A., Lang, S. Q., Lloyd, K. G., Mahmoudi, N., Orsi, W. D., Shah Walter, S. R., Steen, A. D., & Zhao, R. (2020). The fate of organic carbon in marine sediments - New insights from recent data and analysis. *Earth-Science Reviews*, *204*(August 2019), 103146. https://doi.org/10.1016/j.earscirev.2020.103146

Lee, T. R., Wood, W. T., & Phrampus, B. J. (2019). A Machine Learning (kNN) Approach to Predicting Global Seafloor Total Organic Carbon. *Global Biogeochemical Cycles*, *33*(1), 37–46. https://doi.org/10.1029/2018GB005992

Lewis, M., Spiliopoulou, A., & Goldmann, K. (2022). *nestedcv: Nested Cross-Validation with "glmnet" and "caret."*

Masson, D. G., Huvenne, V. A. I., de Stigter, H. C., Arzola, R. G., & LeBas, T. P. (2011). Sedimentary processes in the middle Nazaré Canyon. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *58*(23–24), 2369–2387. https://doi.org/10.1016/J.DSR2.2011.04.003

Masson, D. G., Huvenne, V. A. I., de Stigter, H. C., Wolff, G. A., Kiriakoulakis, K., Arzola, R. G., & Blackbird, S. (2010). Efficient burial of carbon in a submarine canyon. *Geology*, *38*(9), 831–834. https://doi.org/10.1130/G30895.1

Matheron, G. (1965). Les variables régionalisées et leur estimation. Une application de la théorie des fonctions aléatoires aux sciences de la nature. In *Masson et Cie*.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*.

Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210X.13650

Meyer, H., & Pebesma, E. (2022). Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, *13*(1), 2208. https://doi.org/10.1038/s41467-022-29838-9

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling and Software*. https://doi.org/10.1016/j.envsoft.2017.12.001

Mollenhauer, G., Schneider, R. R., Jennerjahn, T., Müller, P. J., & Wefer, G. (2004). Organic carbon accumulation in the South Atlantic Ocean: Its modern, mid-Holocene and last glacial distribution. *Global and Planetary Change*. https://doi.org/10.1016/j.gloplacha.2003.08.002

Montero, J. M., Fernández-Avilés, G., & Mateu, J. (2012). Spatial and Spatio-Temporal Geostatistical Modeling and Kriging. In *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging*. https://doi.org/10.1002/9781118762387

Neto, A. A., Mota, B. B., Belem, A. L., Albuquerque, A. L., & Capilla, R. (2016). Seismic peak amplitude as a predictor of TOC content in shallow marine sediments. *Geo-Marine Letters*. https://doi.org/10.1007/s00367-016-0449-3

Nussbaum, M., Papritz, A., Baltensweiler, A., & Walthert, L. (2014). Estimating soil organic carbon stocks of Swiss forest soils by robust external-drift kriging. *Geoscientific Model Development*. https://doi.org/10.5194/gmd-7-1197-2014

Oliveira, A., Palma, C., & Valença, M. (2011). Heavy metal distribution in surface sediments from the continental shelf adjacent to Nazaré canyon. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *58*(23–24), 2420–2432. https://doi.org/10.1016/j.dsr2.2011.04.006

Oliveira, A., Santos, A. I., Rodrigues, A., & Vitorino, J. (2007). Sedimentary particle distribution and dynamics on the Nazaré canyon system and adjacent shelf (Portugal). *Marine Geology*. https://doi.org/10.1016/j.margeo.2007.04.017

Papritz, A. (2020). *georob: Robust Geostatistical Analysis of Spatial Data. R package version 0.3-13.*

Papritz, A. (2021). *Tutorial and Manual for Geostatistical Analyses with the R package georob*. 1–69.

Paradis, S., Puig, P., Masqué, P., Juan-Diáz, X., Martín, J., & Palanques, A. (2017). Bottom-trawling along submarine canyons impacts deep sedimentary regimes. *Scientific Reports*, *7*. https://doi.org/10.1038/srep43332

Paropkari, A. L., Prakash Babu, C., & Mascarenhas, A. (1992). A critical evaluation of depositional parameters controlling the variability of organic carbon in Arabian Sea sediments. *Marine Geology*. https://doi.org/10.1016/0025-3227(92)90168-H

Payo-Payo, M., Jacinto, R. S., Lastras, G., Rabineau, M., Puig, P., Martín, J., Canals, M., & Sultan, N. (2017). Numerical modeling of bottom trawling-induced sediment transport and accumulation in La Fonera submarine canyon, northwestern Mediterranean Sea. *Marine Geology*, *386*. https://doi.org/10.1016/j.margeo.2017.02.015

Pearman, T. R. R., Robert, K., Callaway, A., Hall, R., lo Iacono, C., & Huvenne, V. A. I. (2020). Improving the predictive capability of benthic species distribution models by incorporating oceanographic data – Towards holistic ecological modelling of a submarine canyon. *Progress in Oceanography*. https://doi.org/10.1016/j.pocean.2020.102338

Pejović, M., Nikolić, M., Heuvelink, G. B. M., Hengl, T., Kilibarda, M., & Bajat, B. (2018). Sparse regression interaction models for spatial prediction of soil properties in 3D. *Computers and Geosciences*. https://doi.org/10.1016/j.cageo.2018.05.008

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., & Pélissier, R. (2020). Spatial validation

reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*. https://doi.org/10.1038/s41467-020-18321-y

Poole, M. A., & O'Farrell, P. N. (1971). The Assumptions of the Linear Regression Model. *Transactions of the Institute of British Geographers*. https://doi.org/10.2307/621706

Premuzic, E. T., Benkovitz, C. M., Gaffney, J. S., & Walsh, J. J. (1982). The nature and distribution of organic matter in the surface sediments of world oceans and seas. *Organic Geochemistry*, *4*(2), 63–77. https://doi.org/10.1016/0146-6380(82)90009-2

Puig, P., Palanques, A., & Martín, J. (2014). Contemporary sediment-transport processes in submarine canyons. *Annual Review of Marine Science*, *6*. https://doi.org/10.1146/annurev-marine-010213-135037

Romankevič, E. A. (1984). *Geochemistry of organic matter in the ocean* (E. A. Romankevich & E. A. Romankevich, Eds.) [Book]. Springer-Verl.

Sala, E., Mayorga, J., Bradley, D., Cabral, R. B., Atwood, T. B., Auber, A., Cheung, W., Costello, C., Ferretti, F., Friedlander, A. M., Gaines, S. D., Garilao, C., Goodell, W., Halpern, B. S., Hinson, A., Kaschner, K., Kesner-Reyes, K., Leprieur, F., McGowan, J., … Lubchenco, J. (2021). Protecting the global ocean for biodiversity, food and climate. *Nature*, *592*(7854), 397–402. https://doi.org/10.1038/s41586-021-03371-z

Schmidt, S., de Stigter, H. C., & van Weering, T. C. E. (2001). Enhanced short-term sediment deposition within the Nazaré Canyon, North-East Atlantic. *Marine Geology*, *173*(1–4), 55–67. https://doi.org/10.1016/S0025-3227(00)00163-8

Seiter, K., Hensen, C., Schröter, J., & Zabel, M. (2004). Organic carbon content in surface sediments - Defining regional provinces. *Deep-Sea Research Part I: Oceanographic Research Papers*, *51*(12), 2001–2026. https://doi.org/10.1016/j.dsr.2004.06.014

Sekulić, A., Kilibarda, M., Heuvelink, G. B. M., Nikolić, M., & Bajat, B. (2020). Random forest spatial interpolation. *Remote Sensing*, *12*(10), 1–29. https://doi.org/10.3390/rs12101687

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. *Proceedings of the 1968 23rd ACM National Conference, ACM 1968*. https://doi.org/10.1145/800186.810616

Sibson, R. (1981). A Brief Description of Natural Neighbour Interpolation. In *Interpreting multivariate data*.

Smeaton, C. (2021). Augmentation of global marine sedimentary carbon storage in the age of plastic. *Limnology and Oceanography Letters*. https://doi.org/10.1002/lol2.10187

Smeaton, C., Hunt, C. A., Turrell, W. R., & Austin, W. E. N. N. (2021). Marine Sedimentary Carbon Stocks of the United Kingdom's Exclusive Economic Zone. *Frontiers in Earth Science*, *0*, 50. https://doi.org/10.3389/feart.2021.593324

Stein, M. L. (1999). *Interpolation of Spatial Data : Some Theory for Kriging.* (1st ed.) [Book]. Springer New York.

Team, R. C. (2021). R: A Language and Environment for Statistical Computing. In *R Foundation for Statistical Computing*.

Thiessen, A. H. (1911). Precipitation Averages for Large Areas. *Monthly Weather Review*. https://doi.org/10.1175/1520-0493(1911)39<1082b:pafla>2.0.co;2

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*. https://doi.org/10.2307/143141

Tyler, P., Amaro, T., Arzola, R., & Cunha, M. R. (2009). Nazaré Submarine Canyon. *Oceanography*, *22*(1), 46–57.

van der Voort, T. S., Blattmann, T. M., Usman, M., Montluçon, D., Loeffler, T., Tavagna, M. L., Gruber, N., & Eglinton, T. I. (2021). MOSAIC (Modern Ocean Sediment Archive and Inventory of Carbon):

A (radio)carbon-centric database for seafloor surficial sediments. *Earth System Science Data*, *13*(5), 2135–2146. https://doi.org/10.5194/essd-13-2135-2021

Vetter, E. W., Smith, C. R., & de Leo, F. C. (2010). Hawaiian hotspots: Enhanced megafaunal abundance and diversity in submarine canyons on the oceanic islands of Hawaii. *Marine Ecology*. https://doi.org/10.1111/j.1439-0485.2009.00351.x

Wadoux, A. M. J.-C., Heuvelink, G. B. M., de Bruin, S., & Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, *457*, 109692. https://doi.org/10.1016/j.ecolmodel.2021.109692

Wainer, J., & Cawley, G. (2021). Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2021.115222

Wakeham, S. G., & Canuel, E. A. (2006). Degradation and preservation of organic matter in marine sediments. In *Handbook of Environmental Chemistry, Volume 2: Reactions and Processes*. https://doi.org/10.1007/698_2_009

Webster, R., & Oliver, M. A. (2008). Geostatistics for Environmental Scientists: Second Edition. In *Geostatistics for Environmental Scientists: Second Edition*. https://doi.org/10.1002/9780470517277

Wessel, P., & Smith, W. H. F. (2017). *A Global Self-consistent, Hierarchical, High-resolution Geography Database (GSHHG)*. http://www.soest.hawaii.edu/pwessel/gshhg/

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v077.i01

Xu, J. P. (2011). Measuring currents in submarine canyons: Technological and scientific progress in the past 30 years. *Geosphere*. https://doi.org/10.1130/GES00640.1

Code

A.1 Computation of virtual sample section depths and derivation of aggregated mean TOC content of a station

```
library(tidyverse)
setwd("~/Documents/ESS/ESSMaster/MA_Thesis/TOC_and_covariates")

# load MOSAIC files
locations <- read.csv("nazare_canyon_cores.csv", sep = ",",
                      stringsAsFactors = FALSE)
core_analyses <- read.csv("nazare_canyon_core_analyses.csv",
                          sep = ",", stringsAsFactors = FALSE)
sample_analyses <- read.csv("nazare_canyon_sample_analyses.csv",
                            sep = ",", stringsAsFactors = FALSE)
#choice of preselection mode
#1: selects samples with average_cm <=h
#2: selects samples by lower_cm <=h (after computation of virtual section depths)

# how should toc_station value be computated
#1: simple mean TOC
#2: weighted mean TOC (corrected version)
#3: ancient algo (nor correct I think)

sel<- 2 # type 1 or 2 (selection of samples)
algo <- 1 #type 1,2 or 3
h <- 2 #chosen depth horizon (for "surficial")

#selection of samples
if(sel==1){ #algo 1 is executed: elects samples with average_cm <=h
  print("selection 1")

  # get subset of cores that have a TOC value and average_cm smaller than maximum surficial sediment depth
  subset1 <- filter(sample_analyses, average_cm <=h & toc !="")

  #get numnber of to be expected unique cores/stations
  unique_cores <- length(unique(subset1$core_id))

  #group by sample_id
  #if there is a replicate value: take the mean value of the replicates (same sample ID=same section depth)
  subset2 <- subset1 %>%
    group_by(sample_id) %>% # group by sample ID
    mutate(mean_repl=case_when(n()>1~sum(toc)/n(), n()<=1~toc)) #adds new column/attribute "mean_repl"

  #get rid of sample replicates
  no_replicates <- filter(subset2, replicate == 1)

  #calculate average sample increments (to derive virtual section limits later)
  no_replicates1 <- no_replicates %>%
    group_by(core_id) %>% # group by core ID
    arrange(average_cm) %>%
    mutate(space = average_cm-lag(average_cm)) %>%
    mutate(total_space=sum(space, na.rm=TRUE))

  no_replicates2 <- no_replicates1 %>%
    group_by(core_id) %>% # group by core ID
    mutate(samples_per_core=n())

  #add new column to df with mean sampling increments
  no_replicates2$mean_space=no_replicates2$total_space/no_replicates2$samples_per_core #lags differ -1 from numbers of core

```

```
  #flag the cores/samples that had been given a virtual upper and lower section depth
  no_replicates2 <- no_replicates2 %>% #  can be twice the same name
    mutate(virtual_horizons =if_else(is.na(upper_cm), "virtual", "original"))

  virtuals <- length(which(no_replicates2$virtual_horizons == "virtual"))

  #if only average_cm given, upper and lower will be derived
  no_replicates2$upper_cm = ifelse(is.na(no_replicates2$upper_cm) & (no_replicates2$average_cm -
                                                                       no_replicates2$mean_space >=
  0) ,
                                   no_replicates2$average_cm - no_replicates2$mean_space,
                                   ifelse(is.na(no_replicates2$lower_cm) & (no_replicates2$average_cm
  -
                                                                             no_replicates2$mean_spa
  ce < 0), 0, #this prevents negative values
                                          no_replicates2$upper_cm))

  no_replicates2$lower_cm=ifelse(is.na(no_replicates2$lower_cm),
                                 no_replicates2$average_cm + no_replicates2$mean_space,
                                 no_replicates2$lower_cm)

  #add section depth magnitude to dataframe
  no_replicates2$depth_fac=no_replicates2$lower_cm-no_replicates2$upper_cm

  selecao <- no_replicates2

} else {
  print("selection 2") # I chose algo 2 : selects samples by lower_cm <=h (after computation of
  virtual section depths)

  subset1 <- filter(sample_analyses, toc!="") # samples with a non-empty TOC attribute

  #get numnber of to be expected unique cores
  unique_cores <- length(unique(subset1$core_id))

  #group by sample_id
  #if there is a replicate value: take the mean
  #value of the replicates (same sample ID=same section depth)
  subset2 <- subset1 %>%
    group_by(sample_id) %>% # group by sample ID
    mutate(mean_repl=case_when(n()>1~sum(toc)/n(), n()<=1~toc))

  #get rid of sample replicates
  no_replicates <- filter(subset2, replicate == 1)

  #calculate average sample increments (to derive virtual section limits later)
  no_replicates1 <- no_replicates %>% # needs to be stored in new dataframe otherwise doesn't work
  (no clue why)
    group_by(core_id) %>% # group by core ID
    arrange(average_cm) %>%
    mutate(space = average_cm-lag(average_cm)) %>%# works
    mutate(total_space=sum(space, na.rm=TRUE))

  no_replicates2 <- no_replicates1 %>%
    group_by(core_id) %>% # group by core ID
    mutate(samples_per_core=n())

  #add new column to df with mean space
  no_replicates2$mean_space=no_replicates2$total_space/no_replicates2$samples_per_core #lags differ -
  1 from numbers of core

  #flag the cores/samples that had been given a virtual upper and lower section depth
  no_replicates2 <- no_replicates2 %>% #  can be twice the same name
    mutate(virtual_horizons =if_else(is.na(upper_cm), "virtual", "original"))
```

```
115.
116.  virtuals <- length(which(no_replicates2$virtual_horizons == "virtual")) #no uf virtual limit
    samples
117.
118.  #if only average_cm given, upper_cm and lower_cm will be derived
119.  no_replicates2$upper_cm = ifelse(is.na(no_replicates2$upper_cm) & (no_replicates2$average_cm -
120.                                                                     no_replicates2$mean_space >=
    0) ,
121.                                   no_replicates2$average_cm - no_replicates2$mean_space,
122.                                   ifelse(is.na(no_replicates2$lower_cm) & (no_replicates2$average_cm
    -
123.                                                                              no_replicates2$mean_spa
    ce < 0), 0, #this prevents negative values
124.                                          no_replicates2$upper_cm))
125.
126.  no_replicates2$lower_cm=ifelse(is.na(no_replicates2$lower_cm),
127.                                 no_replicates2$average_cm + no_replicates2$mean_space,
128.                                 no_replicates2$lower_cm)
129.
130.  #add section depth magnitude to dataframe
131.  no_replicates2$depth_fac=no_replicates2$lower_cm-no_replicates2$upper_cm
132.
133.  # get subset of cores that have TOC attribute and have lower_cm smaller than maximum surficial
    sediment depth
134.  selecao <- filter(no_replicates2, lower_cm <=h)
135.
136.} #end of sample selection loop
137.
138.#toc_station
139.if (algo==1) {
140.  print("algo1: simple mean")
141.  # simple mean for toc_station
142.  no_replicates3 <- selecao %>%
143.    group_by(core_id) %>% # group by core ID
144.    mutate(toc_station=case_when(samples_per_core==1~toc, #if there is only one sample it just takes
    its TOC value
145.                                 samples_per_core>1~(sum(mean_repl)/samples_per_core)))  %>% #
    #otherwise weighted mean TOC value
146.    mutate(max_toc_stat = max(toc, na.rm=TRUE)) %>% #maximum toc at one station
147.    mutate(mean_average_cm=mean(average_cm))%>%
148.
149.    #control to avoid mean TOC being bigger than biggest toc values of a certain station
150.    mutate(toc_station=case_when(toc_station > max_toc_stat~max_toc_stat, TRUE ~ toc_station))
151.
152.    #check if there is no NAs in toc_station
153.    t <- filter(no_replicates3, is.na(toc_station))
154.    if (length(t[t == TRUE]) > 0) {
155.      print('NAs in toc_station!')
156.      print(length(t))
157.      break
158.    }
159.
160.
161.
162.
163.
164.  } else if (algo==2) {
165.  print("algo2: weighted mean")
166.  # weighted mean station TOC
167.  no_replicates3 <- selecao %>%
168.    group_by(core_id) %>% # group by core ID
169.    mutate(toc_station=case_when(samples_per_core==1~toc, #if there is only one sample it just takes
    its TOC value
```

```
                                  samples_per_core>1~sum(mean_repl*depth_fac)/sum(depth_fac)))  %>% #
  #otherwise weighted mean TOC value
    mutate(max_toc_stat = max(toc, na.rm=TRUE)) %>% #maximum toc at one station
    mutate(mean_average_cm=mean(average_cm))%>%

    #control to avoid mean TOC being bigger than biggest toc values of a certain station
    mutate(toc_station=case_when(toc_station > max_toc_stat~max_toc_stat, TRUE ~ toc_station))
    # mutate(toc_station=case_when(toc_station > max_toc_stat~max_toc_stat, toc_station <=
  max_toc_stat~toc_station))

    #check if there is no NAs in toc_station
    t <- filter(no_replicates3, is.na(toc_station))
    if (length(t[t == TRUE]) > 0) {
      print('NAs in toc_station!')
      print(length(t))
      break
    }

  } else {
  print("algo3: ancient weighted mean (erroneous")
  # weighted mean station TOC
  no_replicates3 <- selecao %>%
    group_by(core_id) %>% # group by core ID
    mutate(toc_station=case_when(samples_per_core==1~toc, #if there is only one sample it just takes
  its TOC value
                                  samples_per_core>1~sum(mean_repl*depth_fac)/samples_per_core))  %>%
  # #otherwise weighted mean TOC value
    mutate(max_toc_stat = max(toc, na.rm=TRUE)) %>% #maximum toc at one station
    mutate(mean_average_cm=mean(average_cm))%>%

    #control to avoid mean TOC being bigger than biggest toc values of a certain station
    mutate(toc_station=case_when(toc_station > max_toc_stat~max_toc_stat, TRUE ~ toc_station))

    #check if there is no NAs in toc_station
    t <- filter(no_replicates3, is.na(toc_station))
    if (length(t[t == TRUE]) > 0) {
      print('NAs in toc_station!')
      print(length(t))
      break
    }
} #end of toc_station computation loop


#only keep one per core!!/station!
no_replicates4 <- no_replicates3[!duplicated(no_replicates3$core_id), ]

#check if we get the to be expected amount of cores in the end
if (nrow(no_replicates4)!=unique_cores) {
  print('number of cores in output file differs from expected unique cores')
  break
}

#add row id
no_replicates4$ID <- seq.int(nrow(no_replicates4))

#join the dataframe with the locations dataframe for geomorph setting and sampling method
no_replicates4.1 <- left_join(no_replicates4, locations, by="core_id")

# # exclude columns not needed as otherwise problems with shapefile creation
no_replicates5<- no_replicates4.1[, c('ID','latitude.x', 'longitude.x','depth_m',
                                      'core_id','toc_station','depth_fac','mean_repl',
                                      'mean_gs', 'material_analyzed_x', 'mean_average_cm',
                                      'average_cm','upper_cm', 'lower_cm',
                                      'virtual_horizons', 'samples_per_core','sampling_method_type',
```

```
                                    'geomorphological_site.x')]
#to compare different selection methods and toc computations later
stringname <- paste("h", toString(h), "_sel", toString(sel),"_",
                    "algo",toString(algo), collapse = NULL, sep="") #for later when comparing
    different algo outputs
x <- toString(stringname)
assign(toString(x),no_replicates5)
meantoc <- mean(no_replicates5$toc_station) #mean toc_station
meantocstation <- paste("h", toString(h), "_sel", toString(sel),"_", "algo",toString(algo),"_",
                        "avg_toc_st",sep="",
                        collapse=NULL) #for later when comparing different algo outputs
y <-toString(meantocstation)
assign(toString(y),meantoc)

#write csv file to be opened in a GIS interface
write.csv(no_replicates5,
    paste0("/Users/aline_w/Dropbox/file_swap_thesis/input/cleaned_data_horizon",h,
                                   "_sel", sel,"_algo", algo, "_", nrow(no_replicates4),".csv"),
    row.names=FALSE)

#graphical overview stats
plotty <- ggplot(no_replicates3, aes(x=toc_station)) +
  geom_histogram(bins=30, aes(y=..density..), color="#e9ecef", alpha=0.6, position = 'identity') +
  geom_density(alpha=.2,linetype="dashed") +
  geom_vline(xintercept = mean(no_replicates3$toc_station),        # Add line for mean
             col = "red",
             lwd = 1) +
  annotate("text",                        # Add text for mean
           x =  2,
           y = 2,
           label = paste("Mean =", round(mean(no_replicates3$toc_station),3)),
           col = "red",
           size = 3) +
  xlab("TOC [wt%]") +
  ggtitle(paste0("Histogram and Density Plot for surficial sediment up to ", h, " cm depth"))+
  geom_vline(xintercept = median(no_replicates3$toc_station),      # Add line for median
             col = "blue",
             lwd = 1) +
  annotate("text",                        # Add text for median
           x = 2,
           y = 1,
           label = paste("Median =", median(no_replicates3$toc_station)),
           col = "blue",
           size = 3)
plotty

#print overview: what I've done
print(paste0("hurray, data is clean!", " sample selection method ", sel,
             " has aggregated ", nrow(subset), " samples into ",
             nrow(no_replicates4), " unique stations. toc_station has been derived by algo ",
             algo, ". ", virtuals,
             " samples have computed section limits.", "mean toc_station is ",
             mean(no_replicates4$toc_station), " [wt %]"))
```

A.2 Equidistant custom projected coordinate system

```
1. distances_around_nazare
2. [1] "PROJCRS[\"distances_around_nazare\",\n    BASEGEOGCRS[\"WGS
   84\",\n        DATUM[\"World Geodetic System 1984\",\n            ELLIPSOID[\"WGS
   84\",6378137,298.257223563,\n                LENGTHUNIT[\"metre\",1]],\n            ID[\"E
   PSG\",6326]],\n        PRIMEM[\"Greenwich\",0,\n            ANGLEUNIT[\"Degree\",0.0174532
   925199433]]],\n    CONVERSION[\"unnamed\",\n        METHOD[\"Modified Azimuthal
   Equidistant\",\n            ID[\"EPSG\",9832]],\n        PARAMETER[\"Latitude of natural
   origin\",39.6696944444445,\n            ANGLEUNIT[\"Degree\",0.0174532925199433],\n
        ID[\"EPSG\",8801]],\n        PARAMETER[\"Longitude of natural origin\",-
   10.1085833333333,\n            ANGLEUNIT[\"Degree\",0.0174532925199433],\n            ID[\
   "EPSG\",8802]],\n        PARAMETER[\"False
   easting\",0,\n            LENGTHUNIT[\"metre\",1],\n            ID[\"EPSG\",8806]],\n
      PARAMETER[\"False
   northing\",0,\n            LENGTHUNIT[\"metre\",1],\n            ID[\"EPSG\",8807]]],\n
    CS[Cartesian,2],\n        AXIS[\"(E)\",east,\n            ORDER[1],\n            LENGTHUN
   IT[\"metre\",1,\n                ID[\"EPSG\",9001]]],\n        AXIS[\"(N)\",north,\n
        ORDER[2],\n            LENGTHUNIT[\"metre\",1,\n                ID[\"EPSG\",9001]]]]"
```

Personal declaration: I hereby declare that the submitted Thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the Thesis.