Ca’ Foscari
University
of Venice

Single Cycle Degree programme

In

Data Analytics for
Business and Society

Final Thesis

Supervised and unsupervised assessment of sea level rise risk in coastal zones

Supervisor
Prof. Carlo Giupponi

Graduand
Marco Bidoia
Matriculation Number 866917

Academic Year
2021 / 2022

Index

1. Introduction

1.1. Abstract

The main purpose of this thesis is to quantify the risk related to sea level rise.
This was an empirical exercise which aimed to integrate the fundamental geographical dimension within the classical structure of risk assessments.

The idea behind this work is fundamentally inspired by Tobler's first law of geography[1], which simply states that "everything is related to everything else, but near things are more related than distant things".
While this concept is intuitive, spatial considerations are often neglected in climate change risk analysises involving socio-economic dimensions. At best, this are limited to local assessments.

In parallel with the theoretical efforts on the definition of risk drivers, the analysis of the global distribution of risk is of the utmost importance.
Within this context, the aim is to objectively identify future case-study areas, resulting as highly threateneden from climate change.

To ensure the robustness of the framework proposed, the results of the application of consolidated assessments techniques are compared with the outcomes of novel machine learning methodologies.
The complementariety of these approaches was found as essential for an in-depth comprehension of the phenomenon.

1.2. Climate change and coastal zones

Sea level rise, as a consequence of climate changes, is gaining worldwide attention from governments and also from general public opinion. This may be due to the fact that this phenomenon is already threatening livelihoods and economies of the most exposed communities. The Intergovernmental Panel on Climate Change (IPCC) and its scientific production clearly outline those issues.

In fact, the IPCC, the United Nations body for assessing the science related to climate change and perhaps the most relevant research organization focusing on this topic, has been producing reports presenting increasing evidence on sea level rise since 1990 [1].

In particular, in the sixth assessment report, working group I [2] is explicitly stated that global mean sea level (GMSL)[2] already increased 0.20 cm between 1901 and 2018, with an accelerating rate of rise. Very likely this is the faster rate in the last three thousand years and it's certain that the rise will continue over the 21st century (IPCC, 2021).

[1] Tobler, 1970

[2] Defined in IPCC AR5 (IPCC 2013) as a spatial average of local mean sea levels.

GMSL rise represents a consequence of natural and anthropogenic modifications in the climate system, and even though is the result of complex physical interactions, its main causes could be resumed in 1) warming of the ocean and the related expansion and 2) melting of glaciers and ice sheets (IPCC, 2019).

Ocean expansion is caused by the warming of temperatures which lowers the density of ocean masses, hence, even keeping the mass of the ocean constant, its volume would still be increasing; technically speaking this phenomenon is known as thermal expansion (IPCC, 2019). Higher temperatures are related with an increase in global energy storage and it's now virtually certain that during the latest four decades ocean warming represented the 90% of this increase (IPCC, 2021).

Surely, another fundamental cause of sea level rise is the loss of ice masses. Ice sheets in Greenland and Antarctica, the widest in the world, have the greatest potential contribution to sea level rise, but also glaciers situated in other geographic regions play an important role. To simplify and generalizing, in this case, the physical cause of the increase in sea level is the loss of ice above flotation, adding mass to the ocean. (IPCC, 2019).

Nonetheless, processes related to ice sheet melting, like for instance marine ice cliffs and sheets instability are rather complex and even nowadays are not fully well understood. This is the root of the so called deep uncertainty[3], above all on long term, after 2100, sea level rise projection.

In order to explore this uncertainty, and taking into account the socio-economic variables too, five scenario sets, formally SSPs, have been proposed by IPCC (IPCC, 2021).

They aren't associated with a probability, but are only used for scenario exploration and are meant to represent the whole range of possible future global warming: in the low emission scenario (SSP1) global mean sea level is likely expected to rise of at least 28 cm while in the high emission scenario (SSP5) the increase could exceed 1 m[4] (IPCC, 2021).

Given these premises, it's indeed evident that coastal zones and their communities are severely threatened by climate change effects, even considering sea level rise hazard alone. In fact, coastal risk will "increase by at least one order of magnitude over the 21st century" (IPCC, 2022). Multiple dimensions such as ecosystems, population, livelihoods, infrastructure, food security, cultural and natural heritage and climate mitigation will be involved in those changes (IPCC, 2022). Nonetheless, if current trends are confirmed, population and urbanization are expected to greatly increase in low lying coastal zones, resulting in more than one billion people projected to live in highly exposed location already in 2050 (IPCC, 2022).

Hence, in order to reduce the threats, several actions should be taken, at least on a local scale. Some of them are more "passive" ones, and consist in reducing the

[3] With deep uncertainty I refer to the situation "where analysts do not know, or the parties to a decision cannot agree on, (1) the appropriate conceptual models that describe the relationships among the key driving forces that will shape the long-term future, (2) the probability distributions used to represent uncertainty about key variables and parameters in the mathematical representations of these conceptual models, and/or (3) how to value the desirability of alternative outcomes."(Lempert, 2003).

[4] As compared to 1995-2014 average.

anthropogenic drivers of risk, such as wrong land use planning, unsustainable exploitation of coastal ecosystems or loss of indigenous and local knowledge (IPCC, 2019), while some others are more “active” ones. Those last practises could be regarded as part of the broad concept of adaptation, which I will rigorously define in the following section. Major defense infrastructures, soft engineering, land reclamation and ecosystem conservation and expansion are the main categories of sea level rise adaptation measures (IPCC, 2019). It’s however essential to consider that those measures should be carefully designed, keeping in mind the timing and the planning of the implementation (Haasnoot et al, 2013) and trying to avoid the so called maladaptation, which results in negative outcomes in the long term (Magnan et al, 2016).

On top of that, several factors limit the possibility to cope with sea level rise. Those limits have been classified into soft and hard ones, according to the feasibility of the adaptation options needed to overcome them (IPCC 2022).

Soft limits are mostly related to socio-economic conditions (IPCC 2022). To simplify, facing with the same physical hazard, poor countries are almost for sure expected to be negatively affected far more than rich ones: in the following chapter I will deepen the related concept of adaptation gap and its importance in the context of this analysis will be evident.

Overall, the result which is emerging from the most recent reports on the topic is that any study on the threats posed by sea level rise which aims to be complete cannot neglect the physical or the socio-economic dimensions.
Hard limits are more related to technical factors, and a great reason of concern is the fact that with a global increase in the temperature of 1.5 degrees we will approach several of them (IPCC, 2022): the chance of limiting this increase under 1.5 degrees until 2100 is estimated to be 50% in the best scenario (IPCC, 2021).

1.3. Research questions

The previous introductive section outlined a critical situation: a wide number of lives and assets located in coastal areas are already in a serious risk, and this figure will increase exponentially by the end of the century. Obviously, uncountable scientific efforts, encompassing a wide range of disciplines, socio-economics ones included, have been already dedicated to study this theme and feasible adaptation strategies.

The objective of this thesis is not to try to improve this vast literature. Instead, this an attempt to circumscribe the spatial scope of future researches and specifically of those employing already established modelling frameworks. Classifying the coastal regions in order to identify the areas that are more threatened by sea level rise, as resulting from the combination of vulnerability and exposure, could be indeed of the utmost relevance to the comprehension and further modeling of climate change implications.

However, the global mapping effort underlying this approach remains, to my best knowledge, an open issue. This is true even if the necessity of assessing sea level rise risk on a global scale has been highlighted, at least, already in Gornitz, 1991.

While a growing body of literature focuses on modelling SLR effects at a local or regional level. it’s noteworthy that some attempts have already been made to broaden the scope of the analysis.

On the theoretical side, the IPCC Special Report on the Ocean and Cryosphere in a Changing Climate (IPCC, 2019) and Magnan et al, 2022 deliver important contributions in terms of identification of variables and indicators to measure several dimensions associated with coastal systems risk. They also define and study archetypes of coastal settlements threatened by climate change.
These efforts are indeed precious to drive my work, but significantly differed from the outcome I aimed to obtain with my research. They rely on their domain knowledge and empirical observations on some specific case studies to generalize some theoretical conclusions, while I wanted to apply, to some extent, their results to produce global categorizations of coastal zones.

The DINAS-COAST project and the resulting Dynamic Interactive Vulnerability Assessment model (DIVA) dataset (Vafeidis et al, 2008) represented an attempt to map the coastal zones in agreement with the objectives of this thesis. They proceeded with the subdivision of the global coastlines in homogeneous segments in terms of coastal geomorphology, population density and administrative boundaries. To enable further analyses, they then attributed different vulnerability and impact indicators to each of these segments. Notwithstanding, what is still missing is the classification effort: the authors ensure that the derived coastal units are homogeneous on the segmentation variables, but a comparison between those units is lacking (for example, there is neither a measure of vulnerability nor a grouping of the areas according to their characteristics).
Eventually, even if DINAS-COAST dataset could have been a useful source for carrying out our work, the fact that this projected ended in 2004 and actually the dataset is not publicly available due to maintenance issues forced me to discard the possibility of its exploitation.

At the same time, as already mentioned, other works attempt to classify coastal areas at a regional scale, using indicators derived with different aggregation methods and considering different components of risk (Clemente et al, 2022; Wu, 2021; Tanim et al, 2022; Dossou et al, 2021; Satta et al, 2017, just to cite some of them).
In the existing literature, the most established evaluation frameworks rely on supervised techniques but, more recently, some authors are starting to make use of unsupervised methods.
As defined in Ghahramani, 2004 in “unsupervised learning the machine simply receives inputs but obtains neither supervised target outputs, nor rewards from its environment. By converse, the definition of supervised method could be derived by exclusion from the definition of unsupervised one.
Here, to simplify, any method requiring whatever kind of input other than raw data is defined as supervised.
Both of them have their pros and cons, which will be discussed in section 3.3, hence I believe that comparing the results of both methodologies could be useful to produce a more reliable picture of the situation.
In light of the above, my objective is to merge these approaches, trying to extend the classification effort on a global scale and exploring different, consolidated and novel, methodologies to evaluate the risk posed by sea level rise.

To formalize, the research questions I will try to answer within this contribution are:

- Considering a global scale, which are the sea level rise risk hot-spots? Which are the most threatened territories?

- What are the differences in the results between applying supervised vs unsupervised methods?

- What are the risk dimensions better depicted by each of those approaches?

- Are the methodologies used to derive a global coastal risk index robust broadly speaking?

2. Background

2.1. Coastal zone definition

The first issue I have to tackle before diving into more technical aspects of the thesis is the lack of a commonly shared definition for what is commonly named coastal zone. At a first sight, this fact appears as particularly striking, considering the wide literature on the topic.

Broadly speaking, this definition could be related to distance from the coastline, physical elevation of the area or an ensemble of both of these criteria.
That said, two definitions are recurrent amongst the most frequently cited ones. At the same time those could be able to well represent the object of the following analysis

1. The concept of low elevation coastal zones, as introduced by McGranahan et al, 2007 pointing to "the contiguous area along the coast that is less than 10 metres above sea level". This idea comes back frequently also in IPCC reports (IPCC, 2014 and IPCC,2022 for instance), but embracing this standard alone would mean excluding from the scope of our research several areas that could also be threatened by slr.

2. In fact, other publications on the same topics adopt rougher definitions based on the distance from the coastline. In detail, Lavalle et al, 2011, outlining a standard for EU studies on coastal zones, recognizes that a coastal buffer of 10 kilometers allows "to capture: i) the specific ecosystems and ii) the urban areas that might generate pressure over the coast". I found no reason to believe that this definition applies only to the European context and this is confirmed by Sayre, 2019 who try to identify homogeneous ecological coastal units worldwide.

Nevertheless, I had to recognize that even the latter approach suffered from incompleteness, potentially omitting lands at risk as before. Thus, as I am explaining in the Methods section, I combined these frameworks to geographically circumvent the areas under study.

2.2. Conceptual risk framework

In the following sections I am extensively referring to definitions and frameworks commonly used in climate change literature. It's therefore fundamental to clarify these concepts before diving into our results.

However, it should be noted that, while at a first glance what we are going to describe could appear as sufficiently clear, several conceptual overlaps and "grey-zones" are noticeable going through the vast amount of publications available.

What I aim to establish here is a simple framework useful to ensure that we are on the same page with the reader, systematizing this body of knowledge is clearly behind the scope of this dissertation[5].

Considering the most recent Unep Adaptation Gap Report (UNEP, 2021) as one of the most comprehensive summaries[6] on the topic, here I adopt their definition of risk as "the potential for consequences where something of value is at stake and where the outcome is uncertain, recognizing the diversity of values", as the intersection between vulnerability, exposure and hazard, where

- vulnerability is "the propensity or predisposition to be aversely affected"
- exposure is "the presence of people, livelihoods, species or ecosystems, environmental functions, services, and resources, infrastructure, or economic, social, or cultural assets in places and settings that could be aversely affected"
- hazard is "the potential occurrence of a natural or human induced physical event or trend that may cause loss of life, injury, or other health impacts, as well as damage and loss to property, infrastructure, livelihoods, service provision, ecosystems and environmental resources.

This risk framework was firstly introduced in the IPCC report Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation (IPCC, 2012).

Adaptation is a closely related concept, meaning "the process of adjustment to actual or expected climate and its effects [...] adaptation seeks to moderate or avoid harm or exploit beneficial opportunities." (UNEP 2021)

I want to stress the fact that the latest concept is the more "concrete" of the ones already defined, undoubtedly requiring efforts from the actors involved in the process. One of the primary aims of our effort is indeed to identify areas where adaptation is essential to cope with the risk.

While the concepts of exposure, hazard and adaptation appear as well defined in the literature, the idea of vulnerability seems somewhat vague, in my opinion.

Deepening the question, it becomes evident that vulnerability encompasses a broad range of socio-economic and biophysical dimensions (the latter are sometimes defined as susceptibility, as in Giupponi et al, 2015 for instance).

Adaptive capacity, defined as "the ability of systems, institutions, humans, and other organisms to adjust to potential damage, to take advantage of opportunities, or to

[5] To deepen the topic, see Füssel, 2005, Smit and Wandel, 2006, Giupponi et al, 2015 and IPCC's reports.

[6] It mainly uses different IPCC's reports as sources

respond to consequences” and coping capacity, defined as the “ability of people, organizations, and systems, using available skills and resources, to face and manage adverse conditions, emergencies or disasters” (IPCC,2012), are tied to the concept of vulnerability too.

Hence, as will be presented in the next section, I have chosen to include several measures to account for this complexity.

The close concept of adaptation gap, “the difference between actually implemented adaptation and a societally set goal, determined largely by preferences related to tolerated climate change impacts, and reflecting resource limitations and competing priorities” (UNEP, 2014), emerged as a crucial topic in the most recent literature (IPCC, 2022, for instance). I tried to consider it in this analysis, even if the lack of data explicitly targeted to measure it and the uncertainty stemming from the “subjective” component in its definition constitutes an obstacle to its quantification. Nonetheless, it is reasonable to assume that the variables used to describe adaptive capacity, adaptation and vulnerability in general could serve as a decent proxy for this purpose.

3. Research methodology

3.1. Data processing

Under this section, after presenting the data model used to represent the information on global coastal zones, I will introduce the variables used to map the risk related to sea level rise and, eventually, the choices in terms of analytical methods and models.

3.1.1. Geographical Information Systems: an introduction

In order to achieve the purpose of mapping coastal zones, data should be organized in a specific and well defined spatial format.

The concept of Geographical Information System (GIS) is of the utmost relevance concerning this step of the analysis. This has been rigorously defined as a "combination of hardware, software, human resources and procedures with the purpose of acquiring, managing and analyzing spatial referenced data" (Goodchild and Kemp, 1990).
Nevertheless, in the common language, with the acronym GIS we usually refer to a peculiar kind of spatial data or to a software specialized to deal with those kinds of data.

Most of the works I already cited in section 1.2[7] had to cope with the intrinsic spatial dimension of this kind of analysis even if, depending on the purposes and on the geographic scale, data models adopted could be extremely different.
Just to give a quick overview, GIS data formats could be classified in two main categories: raster data and vector data.

In the raster format, spatial data is represented as a gridded map, subdivided in (usually) equally sized cells[8] or "pixels" associated with the value of the variable of interest. This could be regarded as a two dimensional array in mathematical terms and it is not so different from the common formats used to represent images[9].
In vector format, data could be represented in different kinds of so-called features. The most basic feature type is the point, a simple couple of coordinates; combining pairs of coordinates in more complex formats it is possible to represent other geometric shapes such as lines, polygons and multi-polygons. In order to be used in a meaningful way, those features are usually linked to databases or tables containing the associated attributes, such as the value of the variable of interest.
Common GIS software are designed to handle both formats, and simple algorithms to convert the data from one format to another are usually implemented in those packages.

This is absolutely needed since geographic data, comprised the maps I used and which I am discussing in detail in the appendix, is frequently available only in one of those formats.

[7] Vafeidis et al, 2008; Clemente et al, 2022; Wu, 2021; Tanim et al, 2022; Dossou et al, 2021; Satta et al, 2017

[8] The "size" of the cells is defined as the resolution

[9] I.e..: RGB

However, the data format is not the only source of heterogeneity concerning GIS data: for their spatial representation a coordinate reference system (CRS) is necessary.

In order to associate a measure with a location several steps could be taken, with different options available at each of them. Those operations have been somewhat parametrized, and each of the possible set of parameters is defined as CRS.

To summarize, first of all a choice should be made on the shape of the ellipsoid representing the earth[10], then origin and direction of the axes of coordinates should be defined. This set of information is named the datum[11].

In addition to this operations, it is often needed to project the data from three to two dimensions. Hence, a choice in terms of projection should be made too, considering that there are several trade-offs in terms of preservation of distance, area, shape and direction.

Due to this last alternative, coordinate reference systems are classified in two main classes too: projected and unprojected.
Most of the known CRS are coded according to the European Petroleum Survey Group Standard (EPSG). That's why they are usually identified with an EPSG code.

It is possible to project a map from one CRS to another. Specifically, for raster data formats several algorithms could be employed to complete this task. During the data processing step I employed two main reprojection/resampling algorithms: nearest neighbor and cubic splines.

The first one is the simplest methods and consist in attributing to the reprojected cell the value of the variable associated with the closest cell center in the original map. Despite being rather simple, this method should be employed for spatial data on categorical variables, since it doesn't produce values outside the original set of values, In this analysis, as I will present in the data appendix, it resulted to be effective in the reprojection of continuous data too[12].
The second method is rather complex and consists in fitting piecewise cubic polynomials to the values of the variable in small areas in the original maps and ensuring that those functions are continuous at the border between different zones. The so-computed functions are than used to derive the values in the reprojected map.

For quantitative continuous data, I employed a simple method to evaluate which reprojection method had the best performance in representing each of the variables. I am exposing it in section 3.1.3

[10] The shape of earth, namely geoid, is irregular and therefore it's not directly representable as it is.

[11] Until this point, coordinates are usually defined in degrees (latitude and longitude) and the analyst should consider that distance defined in degree is not constant in terms of meters.

[12] Also for continuos data linked to well defined administraive areas I used only nearest neighbor reprojection in order to avoid smoothing at the borders of this areas and to avoid the production of values different from the original ones

As a last technical note, for GIS analysis I relied on three main open-source solutions:

- Qgis: a software specialized in the analysis of GIS data which comes with a practical GUI and is also customizable with numerous plugins. Here used to perform some of the most advanced and non-automated spatial operations.

- R with its command line accessed through the interface R-studio: a statistical programming language which comes with geo-processing libraries such as terra, extensively used to automate all the workflow when possible.

- Python: a general purpose programming languages, mainly used through Google's Colab service and its notebooks and the Qgis python plugin. It has been employed to perform a small subset of operations such as managing the integration between spatial and non-spatial dataset and customizing Qgis reprojection functions.

3.1.2. Coastal zones data model

As already mentioned, this wide range of alternatives results in many different possible data models to represent coastal zones and sea level rise risk variables.
To cite as example some works which carry out an analysis at least on a regional scale:

- Satta et al, 2017, in the attempt to derive a risk index for Mediterranean coasts, employs high resolution 300m x 300m raster representation of the variables of interest.

- Vafeidis et al, 2008, uses a vector representation of the global coasts: segments of varying length are linked to a database where the values of the variables are recorded. However, the authors report that to elaborate this data model, it has been necessary to elaborate maps of variables with heterogeneous native formats and that the phenomena that have been synthetized often extended far behind the coastline

- Wolff et al, 2018, in an update of Vafeidis et al, 2008 for the Mediterranean region employs the same data model of the previous work, but explicitly admitting that the main disadvantage of raster data model is its computational expensiveness and that the simplified linear data model has been used mostly for this reason.

Theoretically, another alternative using vector format, could have been to employ administrative areas polygons as units of analysis. This was probably a computational efficient solution and could have made easier to link the value of the variables to their location but, at the same time, it could result in a loss of spatial resolution and hence

of information. Some phenomena[13] show indeed a high variability on a scale lower than the administrative district.
On top of that, a choice should have been made on the administrative level to employ, and, even on the same administrative level, there is a huge variability on the average surface area of the administrative areas between the different countries[14].

In this work, I decided to employ a global raster with 1 Km resolution as the baseline data model. This choice has been made to preserve the representation of the spatial variability in the phenomena related to sea level rise risk. In fact, few of the variables I selected are mapped with a native resolution equal or lower than 1km. Despite being more computationally expensive than the linear data model, this resulted in a good trade-off between the simplification in Vafeidis et al, 2008 and Wolff et al, 2018 and the very high resolution in Satta et al, 2017, remarking that this is a global scale analysis.

Another advantage in comparison to the already cited vector data models is the dimension of the sample obtained with this framework. Since I am employing, amongst the others, unsupervised statistical techniques, the sampling of the variables in small pixels instead of the employment of variables associated with administrative areas polygons could result in statistically more reliable estimates of the risk index.

The next data processing issue I had to tackle was the spatial delimitation of the coastal zones. Starting from the theoretical definitions of coastal zone exposed in section 2.1, I had to identify in the maps the borders of the areas which I defined as LEZC and to delimit those regions within 10 km from the coastline.

The first step, common to both operations, was to produce the baseline map of the global coastlines. As a starting point, I used the level zero/national level of the Database of Global Administrative Areas version 4.1 (GADM 4.1). This was the most updated database on state borders at the time of the analysis and comes in vector format.

To select only the coastlines geometries, I derived a single multi-polygon dissolving the borders between states with Qgis native function Dissolve. I then converted this multi-polygon to a linear shape with Qgis native function Polygons to lines.
To select LECZ I applied the procedure described in Merkens et al, 2016:

- employing the CGIAR-CSI SRTM v 4.1 digital elevation model (DEM) and GTOPO30 DEM for high latitudes I was able to classify the areas with an altitude lower than 10 km, with simple raster operations implemented in R terra library

[13] Those variables, presented in the appendix, are the ones that are mapped in a high resolution in the original source, such as land cover or population distribution for instance.
[14] Eg.: the difference in average surface between US states and Italian regions.

- then, I had to ensure the hydrological connectivity with the ocean. With Qgis Gdal function polygonize and 8-connecteness[15] setting I was able to derive polygons of contiguous pixel

- eventually, I intersecated them with the coastlines with Qgis native function extract by location to find the contiguous zones bordering with the ocean and rasterized the file back in the baseline format with terra function rasterize.

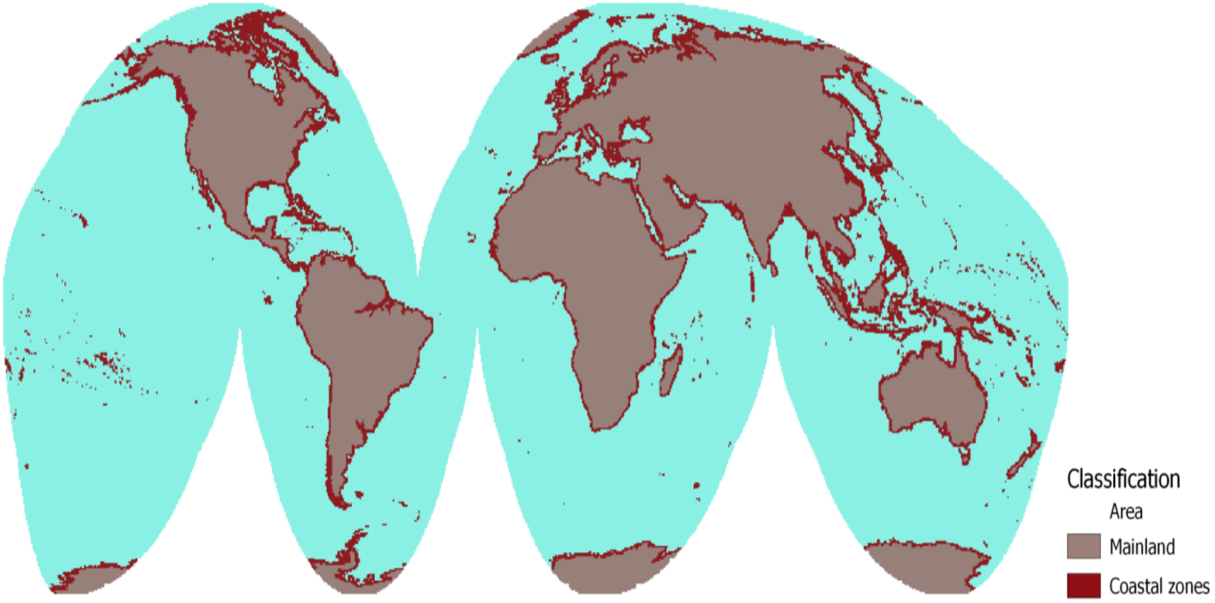
On the other hand, to find the pixels with a maximum linear distance from the coastline equal to 10 km, I applied a simpler procedure:

- I rasterized the polygons created dissolving GADM 4.1 database with terra rasterize, attributing to land and ocean pixels a different code (1/0).

- with Qgis Gdal function proximity I was able to isolate the pixels with the maxima distance of 10 km from the pixels coded as ocean (with values equal to 0).

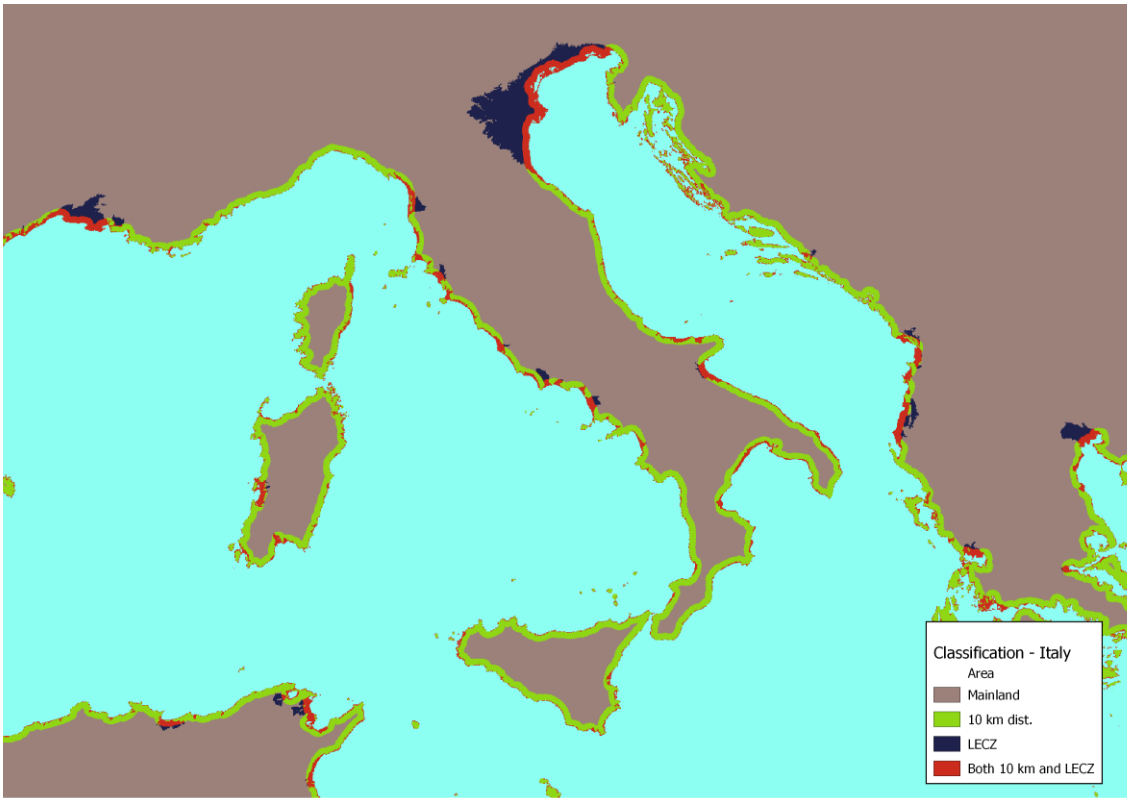
Those informations were then integrated in a single raster file with a mask comprising all the zones here defined as coastal. I maintained a separate code to distinguish LECZ, zones within 10 km from the ocean and zones associated with both definitions. All the other variables are standardized according to this format, and their values are masked to be evaluated only within this spatial range.

Figure 1: Global coastal zones mask



[15] Selects only pixels at an altitude lower than 10 m connected with at least other eight low altitude pixels

Figure 2: Different coastal zones detail, Italy



3.1.3. Projections

Due to the peculiarity of this analysis, I had to made non trivial considerations on the coordinate reference systems to adopt. This the motivation behind the choice of dedicating this whole sections to justify the decisions taken.

The vast majority of the downloaded dataset is associated with a CRS known as World Geodetic System 1984 (WGS84). This is accepted as the international standard (Janssen, 2009), however this is an unprojected CRS, implying that the represented distances in degrees are not constant in terms of meters.

Mainly due to this reason, the employment of this CRS does not make possible to perform several necessary GIS operations in the workflow of this analysis, such as, for instance, the creation of a 10 km buffer around the coastline.

As a rule of thumb, it is always better to introduce minimum distortions by avoiding unnecessary calculations, such as the ones needed to reproject data. However, since this work included raster data from different sources which do not perfectly match in terms of origin, resolution and spatial extent, it is a common practice to standardize all spatial information required as input for a project, within a reference context given by a coordinate reference system, a targeted resolution, a specific origin and a spatial extent. This can be achieved through interpolation or aggregation methods, introducing distortions that become, to some extent, unavoidable for data standardization.

Since in this case it is not possible to work on raw unprojected spatial information and global scale projects typically require processing massive amounts of spatial information, to reduce the computational burden it was highly recommended to employ projected CRS.

A widely used projected CRS is the Universal Transverse Mercator (UTM). This employs of a cylinder tangent to a chosen meridian to project data in two dimensions (Janssen, 2009).

This procedure results in a high distortion moving farther from the reference meridian: that's why, in fact, with the acronym UTM we refer to a "family" of 60 projections with a 6° width, differentiated by their reference meridian (Janssen, 2009).
Hence, conducting this global scale analysis with UTM would imply processing data and running codes separately per each zone, which is not computationally efficient in the case of a global scale analysis

However, several alternative projected CRS covering all the globe exists. Among those, of a particular interest are the equal-area projections. These are designed to preserve the areas, thus representing raster maps with cells of the same size from the equator to the poles.

The reprojection algorithms as a side effect always introduce a distortion that does not allow to represent the true shape of the continents. Nevertheless, the possibility of elaborating informations on cells having the same size can simplify the calculation procedures that necessarily have to consider cells' areal surface.

The equal-area projection which has been shown by Moreira de Sousa et al, 2019 to outclass other equal-area projections in terms of minimization of angular and distance distortions, when only land masses are considered, is the Interrupted Goode Homolosine (IGH).
In its common implementation it makes use of WGS84 datum, which is also important because in this way it is possible to simplify the reprojection operations and hence further reduce distortions.

In contrast with the other CRS presented here, IGH was not associated with an official EPSG code at the time of the anaylsis. However, it is supported by all of the software used in this analysis; it is possible to import it in Qgis with the following Well Known Text[16] (WKT):

[16] A standardized code including the parameters which represent a CRS.

```
PROJCS["Homolosine",
   GEOGCS["WGS 84",
      DATUM["WGS_1984",
         SPHEROID["WGS 84",6378137,298.257223563,
            AUTHORITY["EPSG","7030"]],
  AUTHORITY["EPSG","6326"]],
      PRIMEM["Greenwich",0,
         AUTHORITY["EPSG","8901"]],
      UNIT["degree",0.0174532925199433,
         AUTHORITY["EPSG","9122"]],
      AUTHORITY["EPSG","4326"]],
   PROJECTION["Interrupted_Goode_Homolosine"],
   UNIT["Meter",1]]
```

As already pointed out, the interpolation algorithm that is run while reprojecting maps result in the production of several mismatches between the input and the output map.

While overlaying an unprojected raster map with its projected representation, a mismatch can be detected if reprojection/interpolation displace the underlying projected cell of at least half of the length of the side of the original unprojected cell (about 10 km at equator). This way, sampling projected and unprojected cells with the same sampling grid return different values, i.e. mismatches. This is tipically the case of nearest neighbor algorithm.

Incorrect classification cannot be detected if the projection involves a distortion that moves a cell over an area with identical values[17], but in this case this does not constitute a problem because, at the end, what is important for this analysis is the fact that the original and final map should present the same values at the same locations.
Moreira de Sousa et al, 2019 reports that Mismatches can especially be found at mid-high and high latitudes, but given the level of geometrical complexity of several of the coastlines represented in this analysis I cannot exclude that mismatches could be found on the coastal regions as defined in 3.1.2.

Keeping this in mind, to evaluate the reprojection error and to choose between nearest neighbor and cubic splines interpolation algorithms for non-categorical variables I applied the following procedure:

- I projected the map with nearest neighbor and cubic splines

- I set all the values outside the coastal regions mask equal to null values.

- I generated more than seven millions random points inside the emerged lands with Qgis native function Random points inside polygons, the CRS

[17] e.g. with the same landform or if data are missing

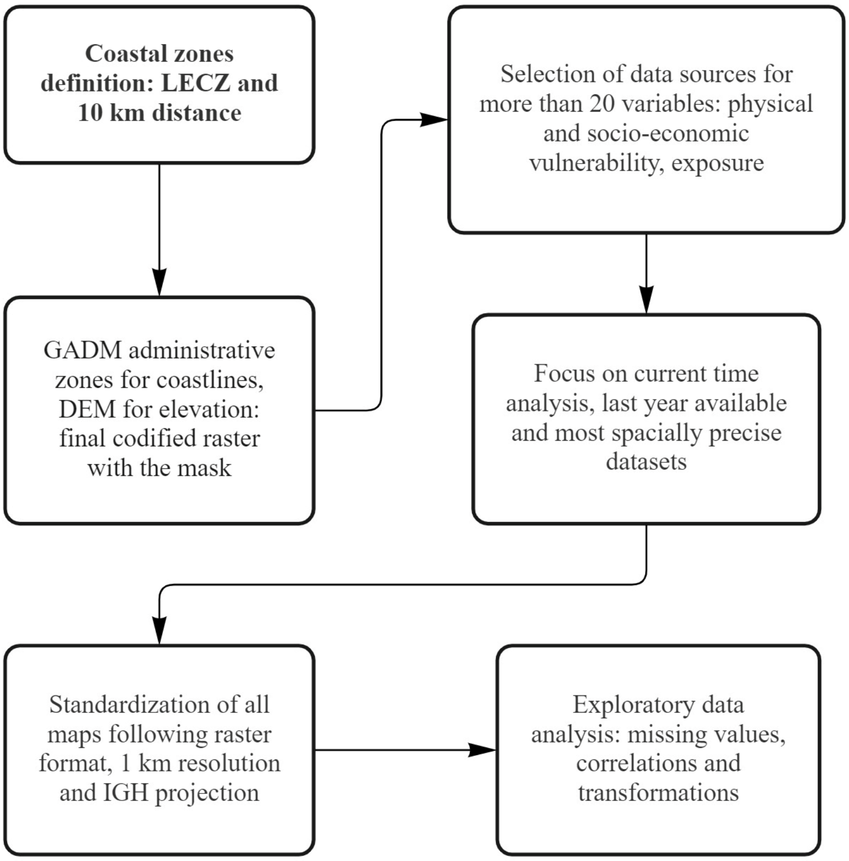
associated was the original CRS associated with each dataset[18]. I saved the points in a vector layes.

- I used the layer file with the points to extract the values of the variable of interest in the unprojected original map and in the projected one, with Qgis native function Raster sampling. In this way I obtained a dataframe in shapefile format with original and final value of the variable at each cell.

- Using this dataframe, the field calculator and the statistics tool in Qgis, I was able to compute error metrics to compare the reprojection algorithms and evaluate the performance of the operations in general. As error metric I decided to use the commonly employed mean absolute error (MAE):

In the data appendix I am presenting the related results, for each of the non-categorical variables used in this analysis.

Figure 3: Data processing simplified flowchart


Coastal zones definition: LECZ and 10 km distance
Selection of data sources for more than 20 variables: physical and socio-economic vulnerability, exposure
GADM administrative zones for coastlines, DEM for elevation: final codified raster with the mask
Focus on current time analysis, last year available and most spacially precise datasets
Standardization of all maps following raster format, 1 km resolution and IGH projection
Exploratory data analysis: missing values, correlations and transformations

[18] In almost all cases the unprojected WGS84 CRS.

3.2. Variables

Here we present the variables we have chosen to map the different components of the framework presented in section 2.2. I am showing the evidence behind my choices, as well as the data sources I selected.

Among the dataset fitting my purposes, I always attempted to select the best ones in terms of resolution, up-to-dateness and consistency.
How to deal with the reciprocal relationship between these variables and with their heterogeneous contributions in terms of relevance to sea level rise risk are issues that are left for the next section.

As a last remark before going through the core of this chapter, I wanted to state that I made the choice to exclusively focus our analysis on vulnerability and exposure of coastal zones to sea level rise. The latter represents the only climate change hazard discussed here, although we should be aware of the other threats to coastal socio-ecosystems. Here, adaptation is considered in terms of measures already taken to reduce exposure and vulnerability.

3.2.1. Vulnerability

As already mentioned, from a theoretical perspective, vulnerability constitutes the variable characterized by the widest dimensionality.
From a biophysical point of view, I accounted for the following factors contributing to vulnerability:

1. Elevation. Naturally, this variable is negatively related to sea level rise and represent a protective factor for the coastal zones. I have included both low elevation coastal zones and zones within 10 km from the coast as object of this analysis, therefore the altitude cannot be assumed as constant. Most of the studies I already cited include this variable, for instance: Wu, 2021, Clemente et al, 2022 and Satta et al, 2017.

2. Distance from rivers. To summarize the motivation behind this choice: "catchment-scale changes have very direct impacts on the coastline, particularly in terms of water and sediment budgets. The changes can be rapid and modify coastlines over short periods of time, outpacing the effects of slr and leading to increased exposure and vulnerability of social-ecological systems" (IPCC, 2019)

3. Presence of coastal ecosystems. Tidal wetlands, coral reefs and seagrasses, contributes to morphological stability and could hence limit the damages of SLR (IPCC, 2019). The conservation and the expansion of these ecosystems could be also part of the adaptation strategies.

4. Coastal erosion. "The material of the coast has a significant impact on the large-scale response to sea-level rise" and "One of the major impacts of sea-level rise is long-term erosion and land loss due to permanent inundation" (Wolff et al, 2018)

5. Anthropogenic subsidence. A factor which significantly worsen the effects of sea level rise (IPCC, 2019, IPCC, 2022). The vertical movement of the land could indeed cause a relative sea level rise locally exceeding the climate related sea level rise (IPCC, 2019)

Including all of the variables related to socio-economic vulnerability appears as a nearly impossible task. Some of these dimensions are intangible and difficult to quantify and map, as for instance the "social values" cited in IPCC, 2019. However, highlighting that adaptive capacity is strictly related to general development issues (UNEP, 2014) I could reasonably be confident that the "classical" set of variables included are able to account for a large part of the variability of this phenomena. The factors considered are:

6. Gross domestic product (Gdp). Availability of economic resources is one of the main drivers of vulnerability, and is also correlated with several other factors contributing to it. "Financial constraints are key determinants of adaptation limits in human and managed systems, particularly in low-income settings" (IPCC, 2022). In this work Gdp is measured in purchasing power parity (PPP).

7. General state of development (Unep, 2014), measured by the Human Development Index, a composite index which takes into account gdp per capita, education and life expectancy and was promoted by the United Nations Development Program (UNDP).

8. Presence of vulnerable social groups. Like children, elderly[19] and women (IPCC, 2022, IPCC, 2019). "Within populations, the poor, women, children, the elderly […] have been especially vulnerable due to a combination of factors" such as, for instance, "gendered division of paid and/ or unpaid labour" and, more in general, health conditions (IPCC, 2022). I measured those groups as a percentage of the total population.

9. Economic inequality. "Societies with high levels of inequity are less resilient to climate change" (IPCC, 2022). This is mainly due to the fact that "SLR and responses may affect communities and society in ways that are not evenly distributed", with also "costs and benefits of action and inaction are distributed unevenly", and this could in turn fuel social conflicts (IPCC, 2019). To measure this phenomena, I employed the Gini index, introduced to quantify di inequality in the distribution of income and widely used in the economic literature (Frank, 2010).

[19] Here, children are < 14 y/old and elderly > 65 y/o

10. Gender inequity. “Gender inequity may be inherent in unfavourable background conditions (higher illiteracy rates, deficiencies in food and calories intake and poorer health conditions) as a result of, among other things, traditions, social norms and patriarchy. Together, these barriers disadvantage women more than men in developing effective responses to anticipate gradual environmental changes such as persistent coastal erosion, flooding and soil salinisation” (IPCC, 2019, IPCC, 2022)

11. Effectiveness of governance

12. Effectiveness of the Juridical system.

    Variables 11. and 12. are frequently cited both in IPCC, 2022 report and in the Unep adaptation gap reports (UNEP 2014 and UNEP, 2021) and , just to summarize their impacts with a few sentences: “prospects for addressing climate-change compounded coastal hazard risk depend on the extent to which societal choices, and associated governance processes and practices, address the drivers and root causes of exposure and social vulnerability” and “concepts of justice, consent and rights-based decision making, together with societal measures of well-being, are increasingly used to legitimate adaptation actions and evaluate the impacts on individuals and ecosystems, diverse communities and across generations” (IPCC, 2022). In this context, those variables are regarded as part of adaptive capacity.

13. Tourism. According to IPCC, 2019, sea level rise can negatively impact landscapes, cultural features and transportation infrastructure and this is going to reduce tourism flows. This is especially true for tourism – dependent areas: I decided to use as a proxy variable the average number of international tourist arrivals (Satta et al, 2017), divided by the national population count

Adaptation measures already taken are generally difficult to measure due to their “active” nature. Finding updated data on adaptation in general is not a trivial task, as highlighted in the Unep Adaptation gap reports too (UNEP, 2014 and UNEP, 2021)

Keeping in mind that I already introduced several variables that could serve as useful proxies to represent adaptive capacity and that complete data on concrete adaptation measures already taken is has been impossible to find, the only adaptation-related variable I considered is only n. 2. presence of coastal ecosystems; which, as previously discussed, could be regarded as both part of vulnerability and adaptation.

3.2.2. Exposure

Compared to vulnerability, determining what variables to choose to depict exposure has been rather easier. Simply going through its definition gives little space for uncertainty in what to include. I took into account:

14. Population count. All the works on the quantification of risk in coastal settings cited so far make use of this variable, which directly comes from the definition of exposure.

15. Percentage of area covered by urban settlements. Again, directly from the definition of exposure

16. Density of main roads and presence of transportation nodes. This is a proxy variable to account for transportation infrastructures in general, and is directly derived from the definition of exposure.

17. Agricultural land cover. As reported in IPCC, 2019, "SLR will affect agriculture mainly through land submergence, soil and fresh groundwater resources salinisation, and land loss due to permanent coastal erosion, with consequences on production, livelihood diversification and food security". As a proxy variable for the degree of dependence on agriculture I use the percentage of the cell covered by agricultural uses.

18. Presence of natural heritage sites. "Coastal risks will increase by at least one order of magnitude over the 21st century due to committed sea-level rise impacting [...] natural heritage" (IPCC, 2022)

3.3. Data analysis

3.3.1. Exploratory data analysis

Before the data analysis in itself, it was necessary to organize the spatial data in a format which enabled a computationally efficient elaboration.

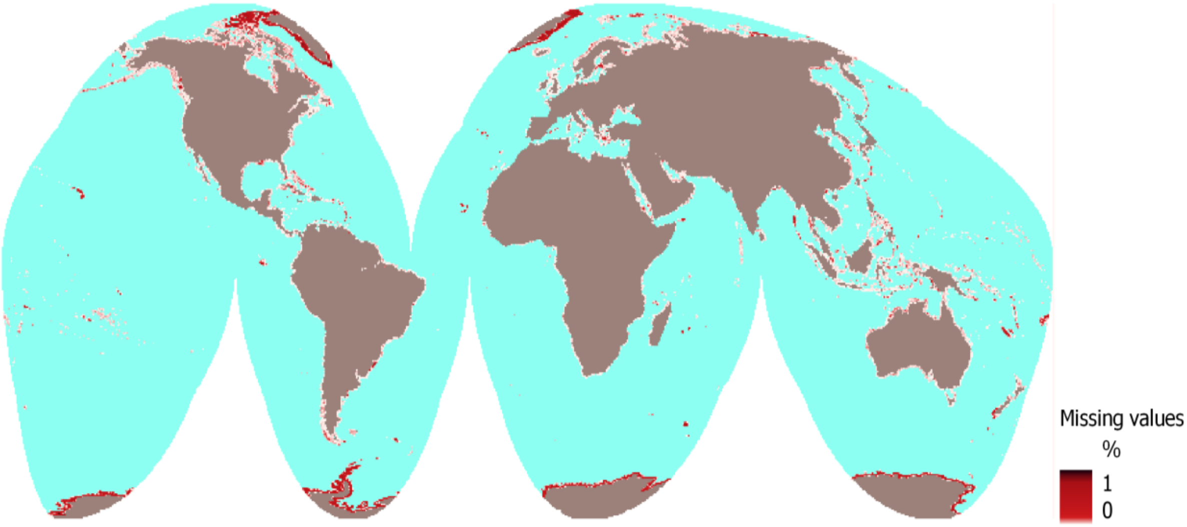
To extract the data, all the raster files containing information on the variables, described in appendix, have been stacked in a unique multiband raster with R raster function: stack.
In this way it was possible to employ the Qgis native function Sample raster values, in order to associate the coordinates of the, previously extracted, centroids of the coastal zones pixels with the values in these locations. Therefore, the data was represented as a collection of points, in vector format.
As a last step of this process, it was possible to dissociate the geometrical components from the other variables in the dataset, retaining the points as simple X and Y coordinate columns and saving the file in the one of the most common formats accepted from programming languages such as Python and R: comma separated value (csv).

This enabled a flexible and efficient use of the powerful statistical and machine learning libraries accepting non-spatial data types as inputs, while maintaining the possibility to convert the data back to spatial formats, for visualization purposes.

This resulted in a dataframe with 7 804 281 observations on 20 variables.

As results from figure 3, the map of points associated with at least one missing value in the dataset, those are mostly distributed in Antarctica and Greenland. Also the coasts near the northern pole and some small islands in the Pacific Ocean the poles present a visually discernible amount of missing values.

Figure 4: Missing values distribution



Going on with the analysis, a first glance to the descriptive statistics in table 1 and 2 intuitively reveals three noteworthy characteristics of the dataset:

- the scale of the variables and their standard deviation in absolute values varies widely across the data, from several orders of magnitude for variables such as population and gdp to a 0-1 range for the ratio of the demographic groups to the total population. The use of scaling procedures is advisable and, for certain steps of the analysis is mandatory, as will be explained.

- Dropping every location point with at least one missing value[20] could still result in retaining 6 509 537 complete observations. The variable prenting the most missing values, anthropogenic subsidence, does not exceed the 1.5 million count of absent data, even including Greenland and Antarctica

[20] Except for subsidence and erodibility, for which some values have been imputed according to the procedures described in the data appendix

- Some variables, such as gdp and roads density for instance, appear as extremely asymmetrical, right skewed, and present a noticeable quantity of outliers. This is confirmed from their histograms and from the related skewness metrics, computed with Python library Scipy. As I am going to explain in the next sections, this could undermine the results of both supervised and, especially, unsupervised methods.

A second fundamental step is the analysis of the structure of correlations between the variables. Given the considerations in the previous points, I decided to employ the Spearman's rank correlation coefficient (Spearman, 1904)

This metric has the fundamental advantages over the standard Pearson correlation coefficient of being robust to outliers and non-linearity in the relationship between the variables: if a monotonic relation exist this tool is more adequate to uncover it.

In practice it is defined similarly to the Pearson correlation coefficient, but instead of immediately employing the value of the variables, those are sorted and ranked and the coefficient is computed on the resulting ranks. If two or more variables share the same rank an average of the ranks of the positions occupied is used (fractional rank).
Defining the rank of a variable as $r(x)$, this results in the following general formula:

$$\frac{COV(r(x_1), r(x_2))}{\sigma(r(x_1))\sigma(r(x_2))} \tag{1}$$

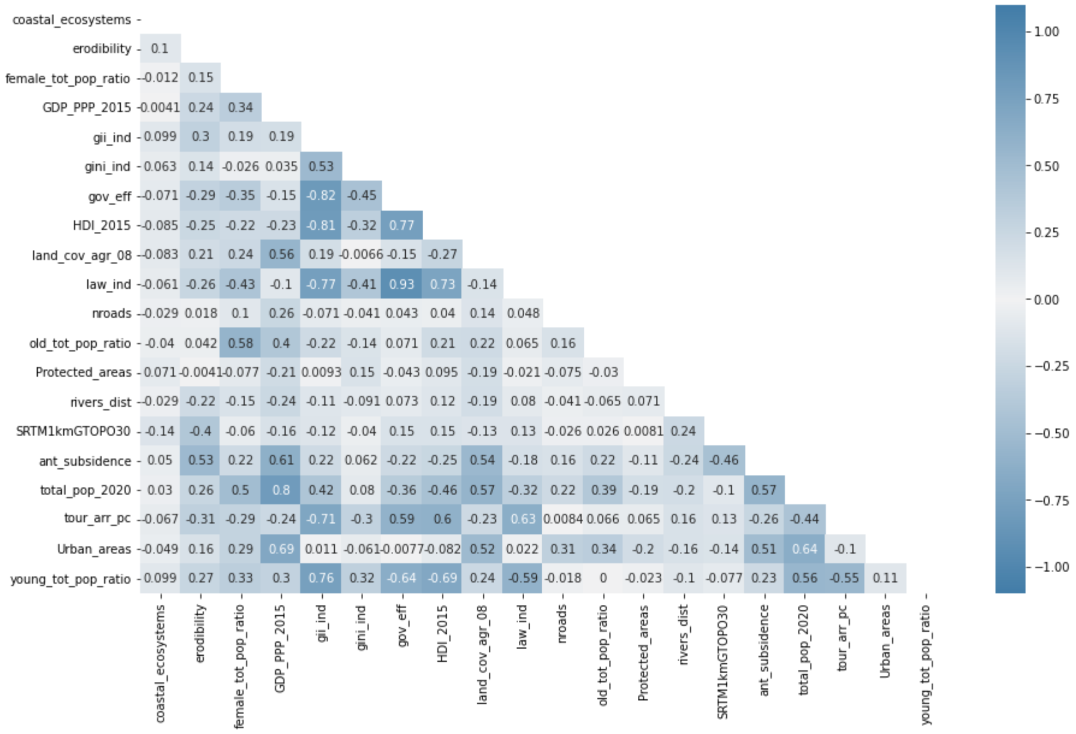
As the Pearson coefficient, Spearman one can take values between -1 and 1, where 1 stands for maxima positive correlation and viceversa.

The features of this coefficient make it particularly suitable for this step of the analysis, where no transformations and standardizations have been applied, making the Pearson coefficient likely to biased given its underlying assumptions. Nevertheless, the results remain valid if only monotonic transformations are employed, due to the use of ranks.

The two variables which are in binary format, presence of protected areas and coastal ecosystems, could also be ranked in terms of impact on risk. Therefore, I could assume that computing the Spearman correlation including those variables respects the assumption underlying the construction of the coefficient and it is well suited for the explorative nature of this first analysis.

The results have been computed with the function spearmanr in Scipy python library, which also includes the p-value related to a test of hypothesis on the significance of the difference of the coefficient from 0. This test is reliable for a sample size greater than 500, as reported in the official references of the function (Kokoska & Zwillinger, 2000).

Figure 5: Correlation heatmap



The correlation heatmap in figure 4 reveals the complexity of the structure of correlations between the variables chosen[21]. Particularly even if unsurprisingly striking is the magnitude of correlations between socio-economic indicators such as law index, government effectiveness, Gini index, gender inequality index, international tourism arrivals and HDI.

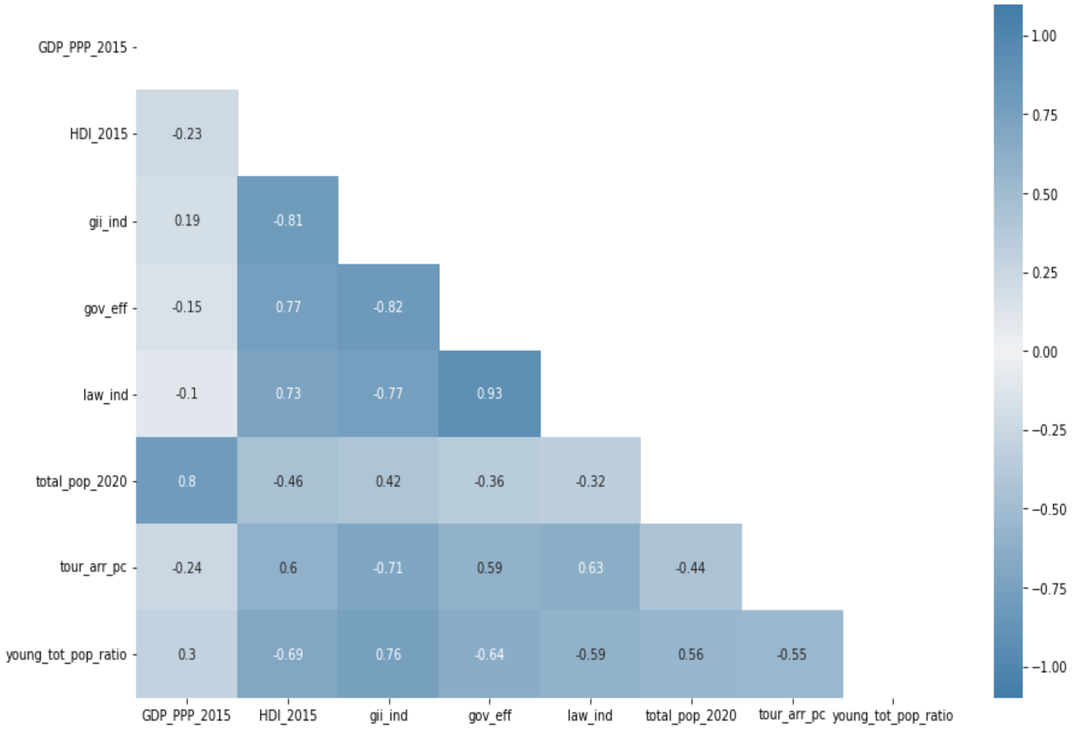
Also total population and Gdp show a high correlation with the other variables in general, it is important to notice that frequently those indicators are used as baseline variable for the spatial projection of other indicator. This remains true even if, as I am going present in the data appendix, when I had to chose the datasets among the ones best suited for this analysis, I attempted to consider sources derived independently one from each other.

Another small group of variables which shows a discernible magnitude of correlations is the one related to subsidence, erodibility and agricultural land cover, which seems to coexist in the same locations.

For an improved graphical focus on those phenomena, I restricted the correlation heatmap to include only the variables with at least one correlation higher than 0.7, figure 5.

[21] The significance of correlations was tested before producing the map, surpisingly only the correlation between old population and young population showed, already lower than 0.001, showed a p-value higher than 0.05 and was then subsituted with exactly 0. In any case there are other correlations that despite being significant are also very low-

Figure 6: Highly correlated variables, heatmap



Given these first results, and without discussing them in detail, it is possible to state that until now the correlations present, broadly speaking, the sign which one could expect from the definition of this variables, without any particular surprise.

In both supervised and unsupervised data analysis approaches, describing and dealing with the relationships between variables is a crucial issue. In the next section I am going to describe the heterogeneous choices I made in this regard.

To conclude this section, as I mentioned before, looking at the descriptive statistics five variables appeared as severely skewed, with a long right tail.
Four out of five skewed variables could be regarded as socio-economic ones: Gdp PPP, density of roads, international tourist arrivals and total population count. Distance from rivers is the fifth variable.
Combining the predominance of socio-economic variables among the skewed ones and also assuming a reasonable non-linear relation of those variables with the sea level rise risk, I decided to employ the simple logarithmic transformation, standard in the econometric literature (Stock & Watson, 2011).

To deal with the presence of zeros in the distribution of all the variables I added a constant c to the variables before applying the transformation. This constant was not fixed for all the variables, but proportional to the minimum non-null value in the distribution[22], following this simple formula:

$$Y = \log(\, X + 0.5 \, \times \min(X \mid X \, > 0)) \tag{2}$$

[22] Those variables always assumed positive values

This choice was made taking into account the heterogeneous orders of magnitude of the asymmetrical variables, ranging from an average value in the order of millions in the case of Gdp to 0.09 in the case of roads density.

This operation was made before the application of other normalization methods, which are going to be different between the two approaches presented in this section.

From the comparison of the histograms before (figure 6) and after (figure 7) the application of formula 2, is already possible to draw different conclusions:

- Concerning the distance from the rivers and the international tourist arrivals, the distribution changes completely, resulting graphically closer to a normal distribution shape.

- Gdp and population are now more symmetrical, even if a strong bimodal tendency is uncovered, with one peak representing the variables which originally were equal to zero or close. Those are gathered in a peak which is somewhat separated from the distribution of the other values. This could reasonably result from the spatial distribution of assets and population, with a non-negligible portion of lands without human presence.

- For the rivers the improvement is moderate: this was originally the second most skewed variable according to the skewness coefficient.

Checking table 2, those first impressions are confirmed, at least concerning symmetry: the only variable left with a fairly high value of skewness post-transformation is the density of roads.

Figure 7: Original variables histogram

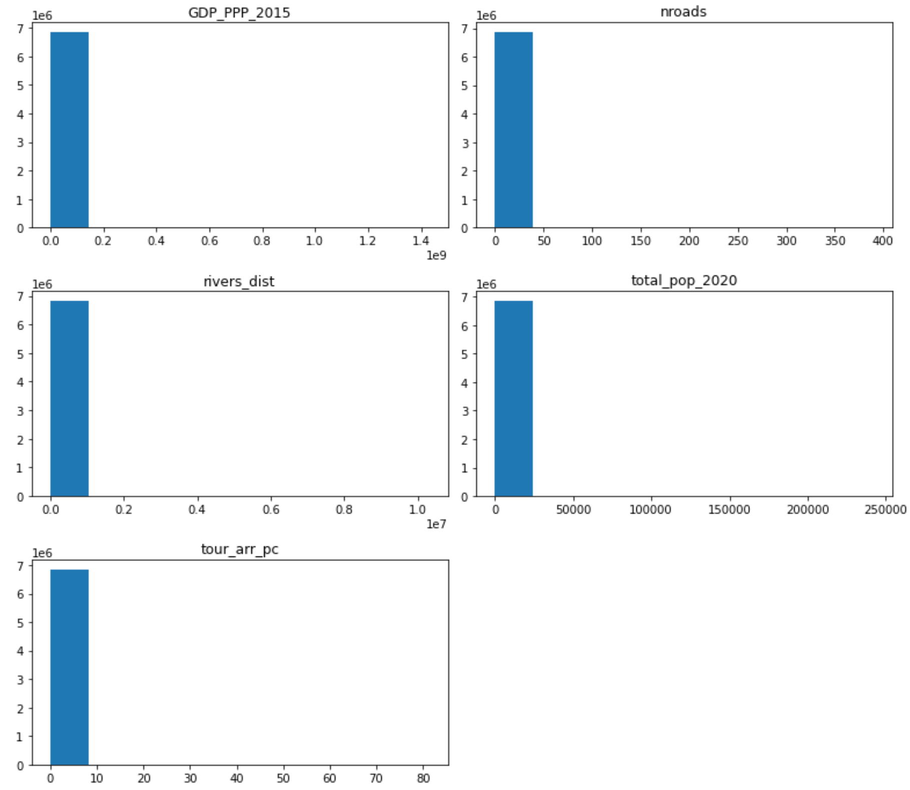


Figure 8: Transformed variables histogram

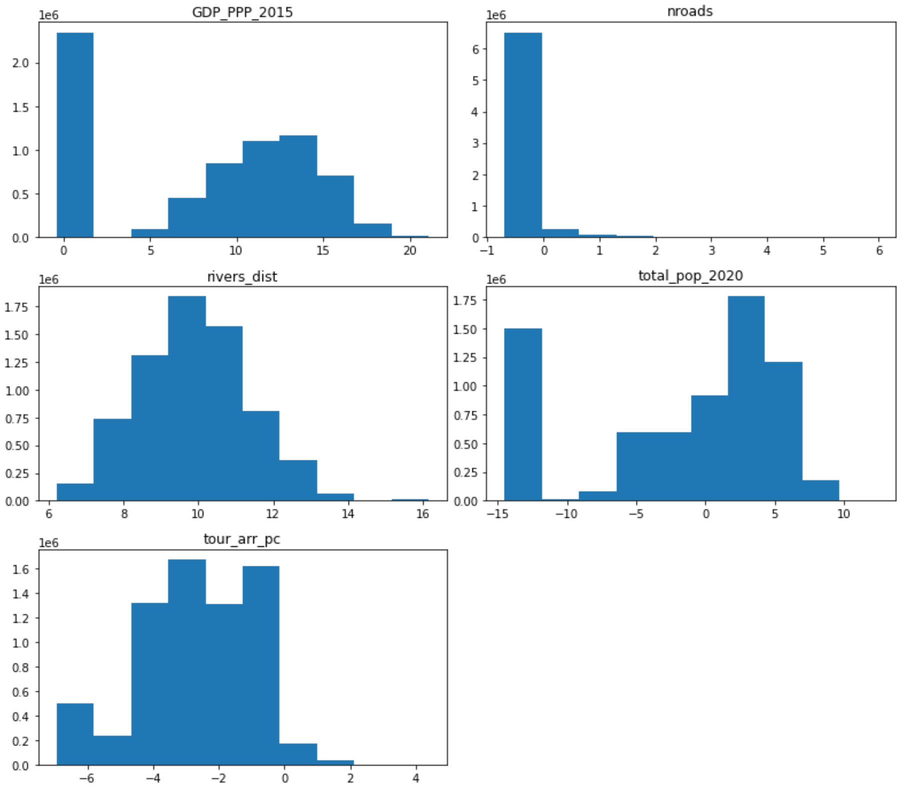


Table 1: Descriptive statistics

	count	mean	std	min	median	max
antr_subsidence	6509537	2,19	1,23	1,00	2,00	6,00
erodibility	7612131	1,81	0,94	1,00	1,40	3,20
female_tot_pop_ratio	6918559	0,38	0,21	0,00	0,49	1,00
GDP_PPP_2015	6968294	2,40E+06	1,91E+07	0,00	1,53E+04	1,43E+09
gii_ind	7803020	0,27	0,16	0,04	0,22	0,80
gini_ind	7803020	37,31	5,54	24,40	36,00	63,00
gov_eff	7803020	0,53	0,89	-2,31	0,37	2,34
HDI_2015	6958285	0,79	0,13	0,32	0,80	1,00
land_cov_agr	7804281	0,18	0,34	0,00	0,00	1,00
law_ind	7803020	0,41	1,06	-2,35	0,24	2,08
nroads	7804281	0,09	0,83	0,00	0,00	390,00
old_tot_pop_ratio	6918561	0,08	0,08	0,00	0,05	1,00
protected_areas	7804281	0,22	0,41	0,00	0,00	1,00
rivers_rast	7804281	3,75E+05	1,36E+06	0,00	2,33E+04	1,08E+07
elevation	7803253	115,84	217,68	-380,99	19,00	3371,00
total_pop_2020	7067033	147,29	947,09	0,00	2,41	2,42E+05
tour_arr_pc	7803020	0,27	0,86	0,00	0,04	81,35
urban_areas	7802941	1,03E+04	4,13E+04	0,00	0,00	9,38E+05
young_tot_pop_ratio	6918559	0,19	0,14	0,00	0,19	1,00

Table 2: Asymmetrical variables: Skewness before and after transformation

Variables	Skw before	Skw after
GDP_PPP_2015	31,89	-0,44
nroads	30,98	5,38
rivers_dist	4,77	-0,08
total_pop_2020	30,96	-0,99
tour_arr_pc	4,86	-0,15

3.3.2. Multiple criteria analysis

What I generally defined as supervised methods, in the sense of requiring inputs from the researcher others than the raw data and the interpretative effort, appear as a consolidated class of methods in the literature of risk assessment (Tanim et al., 2022).

In particular, I am employing Multiple Criteria Analysis (MCA)[23] tools in order to carry out this analysis. This discipline, introduced by Roy, 1968, could be defined as "the use of computational methods that incorporate several criteria and order of preference in evaluating and selecting the best option among many alternatives based on the desired outcome" (Ozsahin et al., 2021).

What emerges from this definition is that the absence of the possibility to objectively rank and weight the different criterias included in the analysis is the main reason behind the necessity to resort to these tools.

These mathematical instruments need also to be contextualized in a more general workflow framework. An optimal framework should also be suited to the objective of the assessment, in this case the analysis of sea level rise risk

Considering those premises, as a general schema for the implementation of this multi-criteria assessment, in the following of this dissertation I am referring to the KULTURisk framework exposed in Giupponi et al., 2015.
The scope of this conceptualization is the "Integrated Risk Assessment of Water-Related Disasters", and I can therefore assume that it is applicable within this context.

As a first step, once selected the variables related to risk in coastal zones as exposed in 3.2, and before starting the analysis in itself, I had to deal with the structure of the correlations between the variables, uncovered in section 3.3.1.

With a customized Python routine, all the pairwise Spearman rank correlation coefficient have been inspected. Between two variables found to have a correlation greater than 0.7 in absolute terms, the one with the greatest average correlation with all the variables in the dataset has been discarded[24].
Given the results of this operation, gender inequality index, law and governance effectiveness index, share of young population and Gdp PPP have been dismissed.

It is important to notice that all of the highly correlated variables belong to the socio-economic categories, in particular the first four variables mentioned appear to be already well represented by HDI and Gini index, while Gdp was correlated with

[23] also defined Multiple Criteria Decision Making (MCDM)

[24] The asymmetrical, variables presented before are employed in their log-transformed form also for this analysis. Given their meaning and the normalization operation that will follow I judged this transformation as reasonable. Otherwise the evaluation could have been strongly influenced by outliers.

several variables, but mostly with population count and urban and agricultural land covers[25].

Following the risk assessment framework, the second step was the normalization of variables.

Given the wide geographical scope of the analysis, the fact that this study covers the vast majority of the global coastal zones, it seems reasonable to state that, for all the variables. values equal or near to both the maximum and minimum extremes are represented in the dataset.

Transforming the variable in a scale ranging from the highest to the lower possible in the world could also have the advantage of being simple to present to the academic audience in the context of the participatory approach that is following.

For metrics contributing positively to risk, a value of one to the maximum and zero to the minimum was assigned and viceversa for the indicators contributing negatively to risk, according to the following standard formula:

$$\begin{cases} z = \dfrac{x - min}{\max - min} \ \ if\ risk\ contr. > 0 \\ z = \dfrac{x - max}{\min - max} \ \ if\ risk\ contr. < \ 0 \end{cases} \tag{3}$$

Normalized the variables and following the assessment protocol, the next operation consisted in the adoption of a participatory approach as weighting procedure. The statistical approach, presented as an alternative in KULTURisk. will be explored under section 3.3.3 – Unsupervised learning.

A questionnaire was distributed to a pool of academic experts on sea level rise and adaptation topics, in an operation of collective elicitation of weights.

Among the six profesionals from which I received an answer, three have an environmental economics background, two an environmental sciences background and one is a researcher in spatial econometrics.

The variety in the profiles of researchers involved in the process should avoid an overall domain biased assignment of weights.

The questionnaire consisted in a simple spreadsheet containing two tables with vulnerability and exposure variables. A brief description of the variables and of the methodologies adopted to derive them was included in the same file.

The experts were required to weight the variables, assigning to them a value comprised between 0 and 100, with the additional constraint that the sum of the weights attributed to the variables belonging to the same group had to be exactly equal to 100.

[25] On one side Gdp contributes to development, but on the other, where the absolute value of Gdp is high there could be also an high exposure of assets and population.

Vulnerability and exposure are treated separately due to the multiplicative procedure of final aggregation chosen, as will be clarified next in this discussion.

In other words, I asked the researchers to express group-relative weights, implicitly forcing them to reflect on the rank of importance of the variables within their groups

The results from this effort are presented in tables 4 and 5 at the end of the section. The set of weights derived by all of the expers are reported along with descriptive statistics such as the mean, the standard deviation and the average rank.
The mean column does not add up to one, since the weights are expressed in relative terms. It is presented only as an immediate summary of the weights' distribution
To complement the drawbacks of the mean in this context, also the average rank is computed. A variable presenting the maximum weights gets a rank value equal to the count of variables in its group (vulnerability or exposure) minus one; at the opposite, the variable with the lower importance gets 0. This value is averaged across all the set of weights.
In the vulnerability group, the highest importance is placed on elevation, according to all the metrics[26], coastal ecosystems presence is stably ranked high, according again to the three statistics.
Tourism arrivals is the vulnerability variable which appears to get less importance, even if showing the highest variance (one researche rassigned to it a very high weight, while the others were negligible).
In the exposure group, presence of urban areas is clearly the most important variable according to all the experts, also population appears as highly weighted, but with a high standard deviation (one researcher assigned a very low weight, while the others were generally high).

The fourth and fundamental step of this process is the employment of an analytical tool to derive the final risk index. To carry out this operation, it is possible to make use of methods ranging from simple averages[27] to more complex algorithms (Giupponi et al., 2015).

Analytical Hierarchical Processing (AHP) (Saaty, 1990), has been already employed in the literature on coastal zones risk (Tanim et al., 2022; Hossain et al., 2022; Le Cozannet et al., 2013).
Starting from weights collected in the same format of this analysis, this method consist in determining the matrix of pairwise, relative, preferences between each couple of variables belonging to the same group.
If a variable X has a value of importance I compared to another variable Z, the importance of variable Z with respect to variable X is assumed to be symmetrical: $1/I$.

The final weights could be computed rowwise or columnwise: the weights associated with each variable are summed and divided by the total sum of weights.

[26] Even if with a fairly high amount of variance

[27] Satta et al, 2017, for instance

Methods based on the eigenvalue/ eigenvector decomposition could be employed to assess the coherence and usefulness of the resulting decision matrix.

In any case, while this method places an emphasis on the description of individual preferences, it is not directly employable to assess the sensitivity of the risk indicator in case of different risk attitudes of the analyst or of the decision maker.
Since MCA tools, and in general risk assessments, are highly dependent on subjectivity, a risk index should not be defined without considering different scenarios of risk attitude. To embedd this considerations I decided to make us of an alternative analytical tool.

Ordered weighted averaging (OWA) has been also already employed on climate change risk assesments literature (Cian et al, 2021; Zhang et al., 2021; Giupponi et al, 2015).

This method has been proposed in Yager, 1988, to balance the aggregation of quantitative criteria between the "orness", the OR condition according to which at least one of the criterias should be satisfied and, at the opposite, and the "andness", the AND condition according to which all the criterias should be satisfied together.

In practice, after a first round of weighting according to the experts' set of weights, the resulting quantitities are sorted and multiplied by a second vector of positional weights.
Those second weights are not related to the importance of the variable in itself but only to its contribution to risk at the level of the precise statistical observation under investigation.

Given n criteria c_i and a first vector of weights w_i, an intermediate set of values v_i is obtained with the simple product $c_i \times w_i$ between those criteria.
The intermediate values v_i are sorted and a last round of weighting with the final weights associated with the position u_p takes place; this is followed by the final aggregation.
In formulas

$$OWA\, risk\, index = \sum_{p=1}^{n} v_p u_p \tag{4}$$

Where v_1 is the highest of the values in the vector of vs and so on.

The second set of weights represents the risk attitude of the decision maker. If those are equally distributed[28] the individual is assumed to be risk neutral. If they increase for the worst performing criterias the individual is assumed to be risk averse, up to

[28] Or high at the extremes and low in the middle

the maximum possible risk aversion if all the weights are set equal to 0 except for the worst performing parameter, which is set to 1.

If they increase for the best performing criterias, the decision maker is assumed to be risk taker, up to the minimum possible risk value if all the weights are set equal to 0 except for the best performing parameter, which is set to 1.

This is the practical implementation of the balancing operation between andness and orness criteria explained before.

It is useful to stress that these operations occur at the level of a single observation / rowwise in a standard dataframe format.

In the context of this analysis, the first set of weights is obtained from the experts' elicitation effort, while three different combinations of order weights are manually selected to represent risk aversion, risk tolerance and risk neutrality.

Plotting the positional sets weights chosen for vulnerability variables in figure 9, with the criteria on the x-axis ordered from the best to the worst performing after the first weighting operation, it is possible to notice that the curve of weights associated with risk aversion is decreasing , the curve of weights associated with risk tolerance is increasing and that the curve associated with risk neutrality has a "parabolic-looking" shape[29]: the worst and the best performing criteria get the same high weights, while the criteria in the middle are associated with lower weights (Cian et al, 2021).

The weighting procedure is executed separately for vulnerability variables and exposure variables, resulting in two different aggregated indexes V and E.

According to the conceptual framework described in Peduzzi et al., 2002 and employed also in Cian et al., 2021, risk is a function of hazard, vulnerability and exposure. Since, in this case sea level rise hazard is kept constant, I employed the following multiplicative function for the final aggregation:

$$risk = f(E, V) = E \times V \tag{5}$$

It seems reasonable to assume that, where there is zero exposure, even in presence of high vulnerability, the risk is still equal to 0 (Peduzzi et al., 2002).

Once made tthese choices, it has been possible to compute the risk values on the whole dataset and for each of the set of weights provided by the experts.

The risk index values obtained from the emloyment of every researcher's set of weight have been again normalized to lie in the range 0-1, where 0 is associated with zero risk and 1 with maximum possible risk in the global coastal zones.

This operation was implemented because the distribution of the different final values resulted often "flattened" on a limited range.

[29] Even if I could have used a flat line as well

In addition, following the same line of reasoning as before, since this is a global map of risk, it makes sense to define the values on a min max scale, where the maximum is the maximum possible value in the world according to each of the experts.
The results from the single experts have then been averaged to obtain a final indicator.

To obtain those results, a customized routine was implemented in Python, where at the time of the analysis it was not possible to find a built-in Owa function.

First of all, an automated data scraping procedure has been implemented to analyze the questionaires, placed in the same folder and with the same naming conventions. Organized the weights from the questionaire, it was then possible to compute the Owa, exploiting the efficience of Python Numpy library array functionalities.

To interpret the indexes it is also possible to employ the quantiles of the distribution of the index. In a map that I will later present in the results section, I selected the top 5% areas associated with the higher risk

Figure 9: Owa weights combination, vulnerability variables

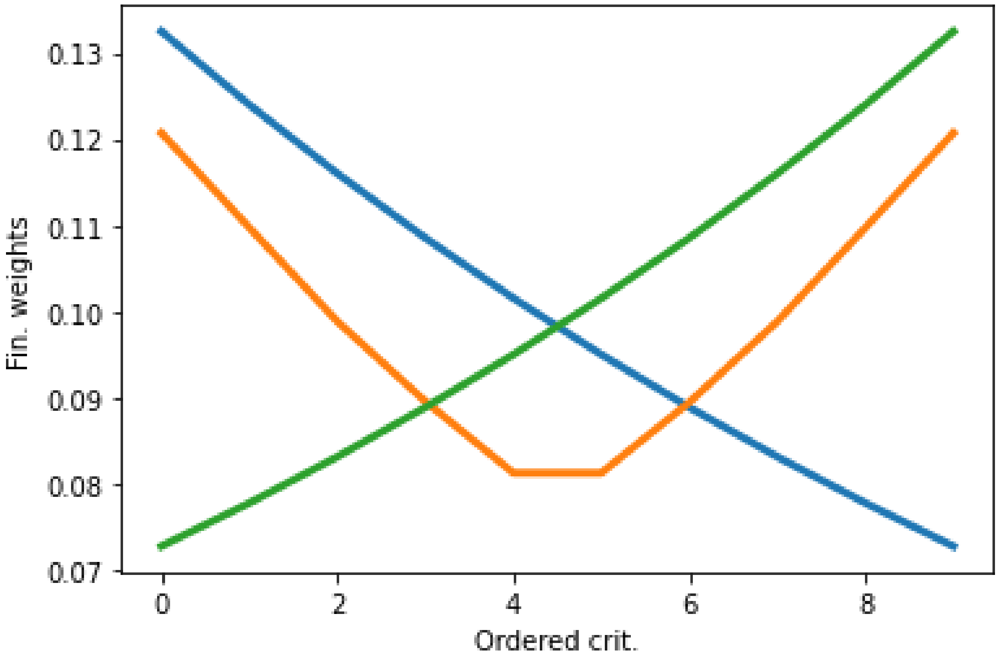


Table 3: Vulnerability weights summary statistics

Variable	E0	E1	E2	E3	E4	E5	Mean	Sd	Av. Rank
ant_subsidence	10	10	6	10	5	10	8.5	2.35	5.3
coastal_ecosystems	10	15	22	10	10	10	12.8	4.92	7.3
erodibility	15	15	1	10	10	20	11.8	6.49	6.8
female_tot_pop_ratio	5	10	3	5	10	3	6.0	3.22	3.8
gini_ind	5	10	15	10	15	10	10.8	3.76	6.3
HDI_2015	10	10	15	12	10	10	11.2	2.04	6.7
old_tot_pop_ratio	5	10	4	7	5	2	5.5	2.74	3.2
rivers_dist	15	10	7	1	15	2.5	8.4	6.02	5.4
elevation	20	5	17	30	15	2.5	14.9	10.10	7.0
tour_arr_pc	5	5	1	5	5	30	8.5	10.65	3.3

Table 4: Exposure weights summary statistics

Variable	E0	E1	E2	E3	E4	E5	Mean	Sd	Av. Rank
land_cov_agr	15	15	17	15	20	15	16.17	2.04	2.33
nroads	10	15	27	15	10	30	17.83	8.61	2.50
protected_areas	20	25	13	10	20	5	15.50	7.45	2.25
total_pop_2020	25	15	1	30	25	20	19.33	10.33	3.17
urban_areas	30	30	33	30	25	30	29.67	2.58	4.75

3.3.3. Unsupervised learning

Data driven methods are increasingly employed in the recent literature on climate change risk evaluation (Tanim et al., 2022).

However, among the available techniques, several still requires a certain degree of "supervision" from the researcher. I will clarify what I mean with this sentence in the following of this section, but I wanted to highlight this concept from the beginning.
In fact, what has driven the choices in terms of unsupervised methods is the idea to limit the intervention of the analyst as far as possible. The idea is to make the comparison with, at the opposite side, a completely supervised method more meaningful.

A brief review on coastal risk quantitative assessment methods present in the already cited Tanim et al., 2022 shows that the majority of studies still employ supervised methods.

Even recent studies applying machine learning techniques to build a coastal or flooding risk index at the end could be classified as supervised, since they make use, usually at a local scale, of data coming from databases related to disasters or damages (Lee et al, 2017; Wang et al, 2020).
For this analysis, being sea level rise, and not risk of extreme events at the coast, the main hazard object of the study, it is also difficult to find validation dataset, above all for a global scale analysis.
Considering the data-driven studies reviewed, another pitfall of those approaches is that they also prevalently take into account physical variables, with an engineering focus.

Studies which come closer to the idea behind this thesis prevalently employ Principal component analysis (Pca) as the most data driven method to construct coastal risk /vulnerability indexes (Tanim et al., 2022; Wu, 2021; Dossou et al., 2021). Methodologies based on this technique are also frequently adopted to evaluate socio-economic vulnerability (Schmidtlein et al, 2008) also in climate risks context (Bucherie et al., 2022).

The reasons behind the need of employing Pca stems from the interrelationship between the different vulnerability and exposure variables. Not only bivariate relationships, but also complex interdependencies driven by latent components which are not directly identifiable in the data.
The idea of this method, introduced independently by Pearson and in Hotelling, 1933 (Mishra, 2017) for psychological analysis, is to identify, given a complex dataset, the best linear combinations of variables which achieve at the same time the result of being uncorrelated between each other and retain the maxima variance of the original data points.
To give a more rigorous mathematical background to this technique, which is at the core of the unsupervised approach I adopted, I refer to Joliffe & Cadima, 2016.

Given the sample covariance matrix S, a vector of weights a, and a data matrix X, any linear combination of the data matrix columns could be described as the matrix product of the weights vector with the data matrix, Xa, and hence the variance of this combination, objective function to maximize is given by the formula:

$$VAR = a'Sa \tag{6}$$

Constraining the optimization to require a set of weights characterized by a unit norm and adopting the Lagrange multipliers procedure, transforms this problem in the study of eigenvalues and eigenvectors of the covariance matrix S:

$$Sa = \lambda a \tag{7}$$

Where a is the eigenvector, the vector with set of weights used to aggregate the original variables in the new component and λ the eigenvalues, the absolute value of the variance explained by the related component.
A number of eigenvectors equal to the matrix columns could be extracted and the sum of all the associated eigenvalues is equal to the total variance.

In this way the data have been decomposed in objects which represent the direction of the correlation between the component, the eigenvectors, and components representing its strength, the eigenvalues.

To obtain the value of the correlation between an original variable and a component it is sufficient to multiply the weight in the vector a_i associated with the nth variable with the squared root of related eigenvector.

$$\lambda_{in} = a_{in}\sqrt{\lambda_{in}} \tag{8}$$

This quantity, L, is defined the as the “loading”, and all the loadings could be organized in the loadings matrix. This matrix is fundamental in the interpretation of the results, used to describe the macro-categories of variables which influence each component.

The loading matrix in its raw version could result difficult to interpret, due to the presence of several non-zero loading for each of the components. One of the goals of the unsupervised learning techniques here employed is indeed the possibility to clearly interpret the results.

This is also a reason of concern for the authors of the papers employing Pca based indicators, that I just mentioned.
Once selected the components of interest and as I am also going to explain in detail, to overcome his issue in all of the aforementioned papers a geometrical transformation of the loading matrix is employed: varimax rotation (Kaiser, 1958).
This algorithm is defined to maximize “the sum of the variances between the squared loadings” (Tanim et al, 2022), in the attempt to attribute to the loading matrix a simpler structure (Kaiser, 1958).

The simpler structure consists in a matrix with few non-zero, high loadings in each column/for each component; these high loadings should preferably occur in different rows for the different columns.
In other word, loadings are “stretched” between a maximum of 1 and minimum of 0. This is done to accentuate the correlation of each component with a small subset of original variables: if one variable, or few variables within the same group, turn out to be strongly associate with one component, it is easier to attribute a meaning to that component in terms of sea level rise risk drivers.

The concept explained here will become more evident when comparing the results in the original loadings table to the ones in the varimax rotated table, generated with R stats function varimax[30] (table 5 & 6)
To close this parenthesis, it is fundamental to mention that this operation modifies the loading table exclusively for interpretative purpose, lefting the structure of the original components unchanged.

All the components resulting from Pca are orthogonal and therefore uncorrelated, it is however important to notice that the objective of this method is to derive the best linear combination of variables according to the variance explained. If the underlying relationships are not linear this could result in a rough simplification of the phenomena.
In any case, given the consolidated employment of this method in the risk indicators literature and the fact that this procedure will be integrated in the analysis workflow with other operations and its results compared with the results from supervised methods, linear approximations seems a reasonable compromise to the use of all the data. A customized selection procedure based on the correlations has been however carried out in the supervised case.

On top of that, other fundamental robustness checks have been carried out to ensure that the dataset is well suited for Pca.

First of all, a simple analysis of correlations with the Pearson coefficient revealed that the strongest correlations found with Spearman rank coefficient are also strongly linear, hence in any case it appears reasonable to synthetize the dataset with Pca.

In second place, the dataset, and specifically its Pearson correlation matrix, has been explicitly tested for Pca suitability with Bartlett test and Kaiser-Meyer-Olkin (KMO) criterion[31], as usually done before the construction of indicators with this method (Wu, 2021; Tanim et al, 2022).

Bartlett test (Bartlett, 1951) has, as underlying null hypothesis, the equality of correlation matrix with identity matrix, implying the absence of correlations. The p-value resulted exactly 0 in this case, meaning that, as observed, there are several correlations in the data.

[30] which also normalizes the loading values by default, as suggested in Kaiser, 1958

KMO (Kaiser, 1970; Kaiser & Rice, 1974) criterion verifies again the presence of linear correlations, checking for the strength of partial correlations. As suggested from the author, a value greater than 0.5 could be considered acceptable. The score obtained with this dataset was equal to 0.81.

Given the linearity assumption and the maximization of variance objective, Pca in itself is not well suited for datasets including dummy variables. A technique has been recently suggested to deal with mixed data types of variables (Pagès, 2004): Factorial Analysis of Mixed Data (FAMD).

Despite this possibility, the scarce documentation of the related R function and the impossibility to find employments of this method in general, and specifically, in the literature on climate risk/vulnerability assessments, drove my choice to select a more consolidated method such as standard Pca. Therefore, the binary variables on protected areas and coastal ecosystems have been discarded from the unsupervised analysis[32].

Before the implementation of the model in itself it was also fundamental to standardize the variables. All of the variables have to be centered at 0 and their variance should be equal, before employing Pca (Joliffe & Cadima, 2016). To achieve this results, I transformed the variables with the classical formula:

$$z = \frac{x - \mu}{\sigma} \tag{9}$$

In practice, this step is common and is already implemented in default in Pca functions in R.

Asimmetry could have been in theory an issue for Pca, which is not limited by the standardization and is not discussed in general on the papers on coastal risk assessment I cited. However, I have already transformed the most skewed variables with the log; only the density of roads remained skewed, and in fact its presence was found to strongly influence the final results: I am going to focus on the results obtained without including that variable, even if I am also going to bring examples of the other case.

Pca has been implemented with R stats function prcomp, the computation of the loadings was done according to expression 5.

Once executed the Pca, it was possible to select the most important principal components[33], as done in all the already cited studies employing the same technique.A number of 5 components, retaining in total slightly more than 70% of the variance appeared as a reasonable choice, also considering at the interpretation of the components and at the graphs.

[32] The variable erodibility presents six ordered categories in terms of risk. I made the choice of keeping it since all the modalities are well represented and spatially distributed (and it is ordered). Pca is also sometimes employed in the literature for the analysis of questionaires with a number of classes comparable to erodibility (Finch, 2017)

[33] Components are ordered according to the variance explained

This threshold is among the most frequently employed ones, even if the rules to follow in this circumstance are associated with a certain degree of discretion (Joliffe & Cadima, 2016).
gh
Including the density of major roads, the change of the results in terms of choice of principal components and general metrics of the model has been negligible: the first five components were again associated with 70% of the variance.

From the raw loading matrix in table 4 it is already possible to identify some patterns in the data, even if the non-negligible amount of fairly high loadings in the entries of the matrix make it harder to uncover the meaning of the components. Moreover, the first component is strongly associated with most of the variables
This is the reason behind my choice to expose the same patterns employing table 3, with the results of Varimax rotation[34].

Now I proceed exposing my observations on every component, taking care of the difference in the variance of the components:

1. The first component appears related to general development issued: the loadings are high in magnitude and positive for government effectiveness index, HDI, law effectiveness index and tourism arrivals, high in magnitude and negative for the share of young population and the gender inequality index. Except for tourism arrivals, this component appears as negative related with socio-economic development variables associated with risk increase and viceversa. Hence, I can assume that its contribution is positive in general[35].

2. The second component is strongly negatively related with Gdp, total population, old population and female population. Even if the relation between sea level rise risk and Gdp should be theoretically negative, its relation with the other variables is clearly positive[36]. Therefore, I would sustain that this component gives a positive contribution to risk reduction, as the first one.

3. The third component appears strongly related with bio-physical aspects of sea level rise risk: it is negatively correlated with altitude and positively correlated with agricultural land cover, subsidence and erodibility. The first variable should in theory reduce the risk while the others should increase it. Therefore, the contribution of this component in terms of risk reduction is negative

4. The fourth component is negatively related with Gini index and positively related with the distance from rivers. While the contribution in terms of risk

[34] In both tables, loadings higher than 0.5 are highlighted in grey. Results are robust to inclusion/exclusion of road density (compare table 6 and 7)
[35] The sign in table 5 near to the column name is the sign of the contribution in terms of risk reduction
[36] One plausible explanation of this phenomena, confirmed also by the simple analysis of correlations, is that, being gdp defined in absolute terms, it is possible that the development variables in component 1 are already sufficient to depict its contribution in terms of vulnerablity reduction: what is "left" uncovered is its relation with the exposure of assets and people, which emerges in the second component.

reduction in this case is undoubtedly positive, giving a precise meaning to this component result more difficult. Comparing the transformed table with the original one, it is possible to notice that in origin this component was only strongly related with Gini index.

5. Fifth component is only strongly and positively related with urban areas (when the density of roads is included, also that variable is related with this component). Hence it is related with an increase in the exposure: the contribution to risk reduction is negative.

Once reached this step, in the works on Pca based risk indicators that I mentioned at the beginning of this chapter, the original variables are weighted according to the results of the Principal components analysis. This is done in heterogeneous ways, or directly employing the components or building intermediate indicators aggregated in subsequent steps (Wu, 2021). A sign to the components is always explicitly attributed. What is noteworthy is that the variance explained by each component usually plays the most important role in this weighting schema, being a fundamental parameter for the final aggregation.

Since here I am already employing supervised techniques to classify risk, another unique indicator is not fundamental for this analysis. In the case of this analysis, as I am explaining, the components will still receive a certain amount of weight in terms of variance they explain.

An alternative approach which could be of interest at this point is the clustering of coastal zones in areas sharing the same characteristics of vulnerability and exposure. This is also complementary to supervised techniques: once individuated the risk hot-spots it should be possible to explore to which cluster the majority of observations in the hot-spot belong to, and, in turn, understand what are the components which of the cluster which are more related to sea level rise risk.

While I was not able to find a paper on coastal risk assessment employing Pca followed by clustering techniques, this workflow is frequently adopted, also in environmental economics and climate change literature (Gonzalez et al, 2022; Alaniz et al, 2022; Addy et al, 2021)-

Several clustering techniques could be employed; among those, popular ones are K-means (MacQueen, 1967), employed in Addy et al, 2021, Hierachical Clustering, and particularly Hierarchical Agglomerative Clustering (HAC) which “has been the dominant approach to constructing embedded classification schemes” (Murtagh & Contreras), used in Alaniz et al, 2022. Another alternative approach which has been explored for this dissertation is the Density Based Spatial Clustering of Applications with Noise (DBSCAN, Ester et al., 1996).

For this short review also I found a precious reference in the Python’s machine learning library SciKit documentation (Pedregosa et al, 2011), which I have also employed to implement the clustering algorithms.

DBSCAN belongs to the density based clustering techniques in the sense that it is designed to find continuous regions of data points in the parameters space. In other words, it aims to find regions where spatial concentrations of points are distinguishable.
To carry out this operation it is necessary to define what is meant by core point in terms number of other points which lies in its proximity and when a point is located in the neighborhood of another point. Point outside any neighborhood are marked as noise and are not assigned to any of the clusters.
The main parameters to be chosen are the number of neighborhood points to be included in a dense region and the maximum distance of two points to be considered near.

HAC is a hierarchical algorithm with a bottom-up approach. It starts from the distance matrix between each pair of observations and aggregates the points in fewer groups at every iteration, until a unique cluster or a number of clusters lower than a selected threshold is formed. This does not produce a single clustering, but, instead, a hierarchical tree of possible segmentations.
The method to derive the distance to be considered in the aggregation has to be selected, and could be, for instance, the minima or the maxima distance between two groups of points. In addition, the hierarchical level or levels of clustering to be employed has to be choosen too[37]

Both those clustering techniques have fundamental advantages over K-Means, such as the robustness to outliers, the fact that they allow for non-spherical shapes of clusters and the fact that clusters could have very different sizes. However, their requirements in terms of computational complexity and memory, combined with the resources available at the time of this analysis and the size of the dataset have been prohibitive for their employment.

Since the data comprises almost all the coastal zones of the world mapped in detail, the aim of this analysis was to describe and classify the data as a whole, and not to compute predictions on the clusters of risk with few observations compared to the size of population.
This is main the motivation behind my choice of exploiting the entire dataset available with a simpler, yet powerful technique, such as K-Means clustering. I am now explaining this method in detail.

As in the case of Pca, data should not be asymmetrical. Employing raw data, another fundamental requirement is the correct normalizaion of data, if what is wanted is that the variables have the same influence on the final results

Being linear combinations of data centred at 0, the principal components already have a 0 mean by definition but their variance is equal to the absolute portion of the variance they retain from the original dataset, as already explained.

In the literature examples making use of K-Means after Pca this fact is never mentioned and the components are directly employed without additional operations[38].

[37] Usually done graphically with the help of a dendogram (Ketchen & Shook, 1996)

[38] Nevertheless, some works have shown that Pca is mathematically related to K-Means clustering (Xu et al, 2015

This remark has been made to notice that I am aware of this phenomena; I have already mentioned that in the classical construction of risk indicators via Pca, the first components usually get more weight: in this analysis this is going to happen without requiring additional assumption from the analyst.

K-Means could be classified as a partitioning clustering technique which, given a number of clusters N, divides the data points p in N clusters C. Each point is assigned to exactly one cluster and the clusters could be represented by their centroids c, the vectors of the average values of the variables in that group.

The objective function to maximize, in the classical version of the method[39], is the sum of squared distance between the centroids and the points (SSD), or inertia:

$$\min(SSD) = \min(\sum_{i=1}^{N} \sum_{p_t \in C} (c_i - p_t)^2) \tag{10}$$

The standard K-Means algorithm starts with the random selection of N points[40] as the first centroids. At every iteration the new centroids c are updated with the averages of the points within the previous cluster.
The default number of iterations in Scikit function is 300 and this process could terminate at a local minimum (Pedregosa et al, 2011).

Usually, the results in terms of SSDs with different number of clusters as inputs are compared. This should be done keeping in mind that this metric mathematically decreases or at least remains equal following the increase in the number of clusters. One simple but frequently employed method to determine the desiderable number of clusters is a graphical one.
It consists in plotting the values of the SSD for the different clustering solutions obtained with an increasing number of clusters, this parameter is then approximately chosen according to point the of the graph in which the magnitude of the decrease of the error function becomes distinctly lower: the location where an “elbow” is formed in the graph.
This is the reason why this approach is defined as the elbow method. (Ketchen & Shook, 1996).

As it is possible to imagine, this process is highly discretional, driven by partially subjective impressions.
After this graphical analysis three possible Ns have been selected for further inspections, also considering the presence tradeoff between the interpretability of the clustering resuls and the number of clusters: 5, 7 and 10[41].

[39] I used the Scikit default euclidean distance

[40] A slighty modified version of the algorithm, named K-Means ++ is employed as defaul in SciKit learn. It selects the centroids according to the probability distribution of the data points (Arthur & Vassilvitskii, 2007)

[41] Also the case of N= 6 has been considered, but therelated results will not be presented here since it evidently did not add information as compared to N=5

In fact, as noticeable in tables 7, 8 and 9 at the end of the section and as I am exposing in results section, the interpretability of the outcomes requires an increasing effort as the number of clusters increase.
This happened both for the interpretation of the clusters centroids in terms of values of the principal components and from a geographical point of view in the maps, with an increase in the noise of the spatial segmentation of areas.

The considerations on the spatial distribution of the clusters, as already presented in detail in section 3.3.3. are the ones that have mostly driven my choices in terms of optimal clusters number.
Even if I am presenting results on clustering obtained with N equal to 5, 7 and 10, for more in-depth analyses and comparisons with unsupervised learning results I have chosen the select the lower number of clusters.

To conclude this section, I am now interpeting the cluster centroids in table 8. During the interpretation of the centroids care should be taken in considering that the variance of the components is diffent:

0. Average conditions cluster: the first cluster seems associated with near to average values for all the components. It is only mildly negatively related to absence of population exposure and vulnerable demographic group.

1. Developed areas cluster: the second group is positively associated with high values of development variables in the first component, but strongly negatively associated with absence of population exposure and vulnerable demographic group in the second component. It is negatively related with physical vulnerability represented in the third component. The average risk value is expected to result balanced within this group, without exceeding high extremes.

2. Threatened areas cluster: he third cluster is negatively related with all the variables reducing risk and positively correlated with the variables increasing risk. This group should be carefully compared to the high risk areas from the supervised approach

3. Optimal conditions cluster: This group is strogly positively related with development variables and absence of population exposure and vulnerable demographic group, it is mildly positively related with physical vulnerability and presence of urban areas.

4. Least developed areas cluster: this group is strongly negatively associated with development variables, the values of the other variables are close to the average. It could depict strong socio-economic risk but average conditions of exposure and physical vulnerability

Table 4: Pca summary metrics

P.C.	Std. Dev.	Prop. var.	Cum. prop. var.
1	2,470	0,359	0,359
2	1,680	0,166	0,525
3	1,245	0,091	0,616
4	0,973	0,056	0,672
5	0,966	0,055	0,727
6	0,943	0,052	0,779
7	0,881	0,046	0,825
8	0,833	0,041	0,865
9	0,741	0,032	0,898
10	0,686	0,028	0,925
11	0,621	0,023	0,948
12	0,563	0,019	0,967
13	0,484	0,014	0,981
14	0,356	0,007	0,988
15	0,311	0,006	0,994
16	0,254	0,004	0,998
17	0,205	0,002	1,000

Figure 10: Cumulative share of variance explained by principal components

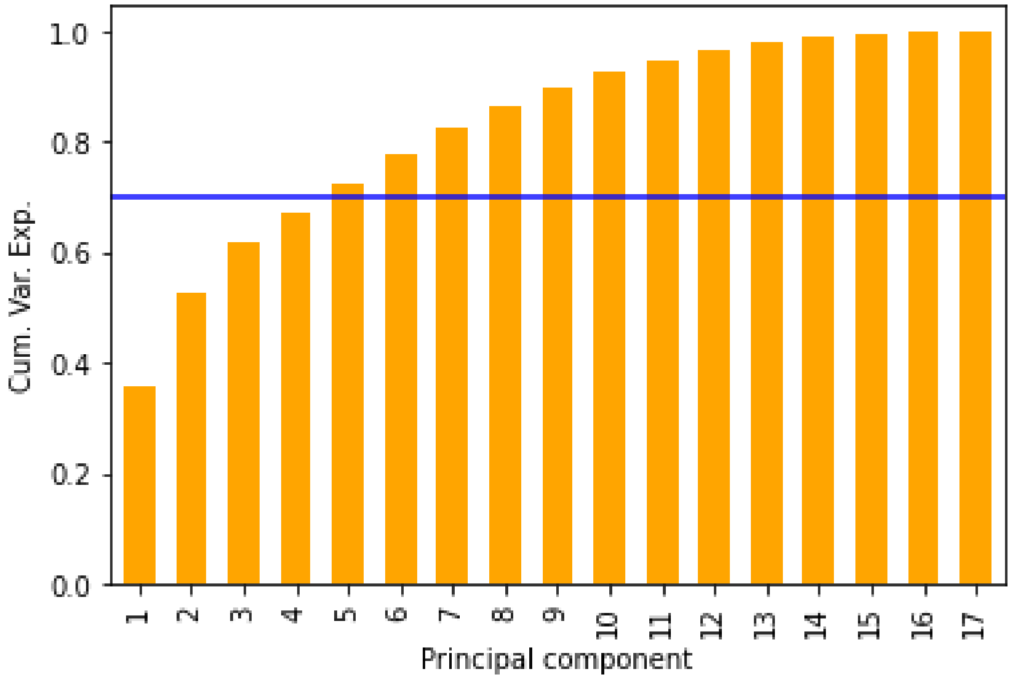


Table 5: Loading matrix

Variable	PC1	PC2	PC3	PC4	PC5
erodibility_index	-0,467	-0,224	0,505	0,103	-0,142
female_tot_pop_ratio	-0,674	-0,373	-0,490	0,033	-0,176
GDP_PPP_2015	-0,571	-0,599	-0,053	-0,135	0,089
gii_ind	-0,833	0,416	0,035	-0,055	0,024
gini_ind	-0,337	0,325	0,096	-0,644	-0,167
gov_eff	0,824	-0,402	0,029	-0,036	-0,064
HDI_2015	0,794	-0,381	-0,039	-0,083	-0,094
land_cov_agr	-0,418	-0,446	0,199	0,305	0,106
law_ind	0,816	-0,367	0,025	-0,064	-0,054
old_tot_pop_ratio	0,058	-0,669	-0,472	0,007	-0,236
rivers_rast_dist	0,219	0,237	-0,333	0,314	0,610
elevation	0,250	0,178	-0,521	-0,190	0,153
subsidence	-0,472	-0,580	0,451	0,067	0,147
total_pop_2020	-0,771	-0,468	-0,331	-0,027	0,018
tour_arr_pc	0,705	-0,212	0,003	-0,075	-0,018
urban_areas	-0,135	-0,453	0,132	-0,497	0,583
young_tot_pop_ratio	-0,837	0,119	-0,250	0,021	-0,074

Table 6: Varimax rotated loading matrix

Variable	PC1 +	PC2 +	PC3 -	PC4 +	PC5 -
erodibility_index	-0,226		0,686	-0,170	
female_tot_pop_ratio	-0,394	-0,839			
GDP_PPP_2015	-0,181	-0,613	0,371		0,404
gii_ind	-0,914		0,103	-0,156	
gini_ind	-0,393	0,119	-0,163	-0,659	0,204
gov_eff	0,913				
HDI_2015	0,878		-0,147		
land_cov_agr	-0,133	-0,306	0,585	0,220	0,122
law_ind	0,888		-0,112		
old_tot_pop_ratio	0,385	-0,761			
rivers_rast_dist		0,166	-0,389	0,700	0,116
elevation			-0,634		
subsidence		-0,244	0,764		0,372
total_pop_2020	-0,435	-0,793	0,220		0,239
tour_arr_pc	0,710	0,125	-0,168		
urban_areas		-0,119	0,112		0,891
young_tot_pop_ratio	-0,779	-0,407			

Table 7: Varimax rotated loading matrix, roads density included

Variable	PC1 +	PC2 +	PC3 -	PC4 +	PC5 -
erodibility_index	-0.218		0.690		-0.182
female_tot_pop_ratio	-0.395	-0.832			
GDP_PPP_2015	-0.184	-0.622	0.404	-0.314	
gii_ind	-0.911		0.111		-0.165
gini_ind	-0.389		-0.139		-0.669
gov_eff	0.914				
HDI_2015	0.876		-0.152		
land_cov_agr	-0.125	-0.318	0.609		0.197
law_ind	0.889		-0.109		
nroads				-0.807	
old_tot_pop_ratio	0.385	-0.755			
rivers_rast_dist		0.151	-0.352		0.701
elevation			-0.599		
subsidence		-0.254	0.799	-0.257	
total_pop_2020	-0.439	-0.794	0.237	-0.207	
tour_arr_pc	0.711	0.119	-0.163		
urban_areas		-0.121	0.167	-0.789	
young_tot_pop_ratio	-0.776	-0.408			

Figure 11: N° of clusters against SSE (in 10 mil)

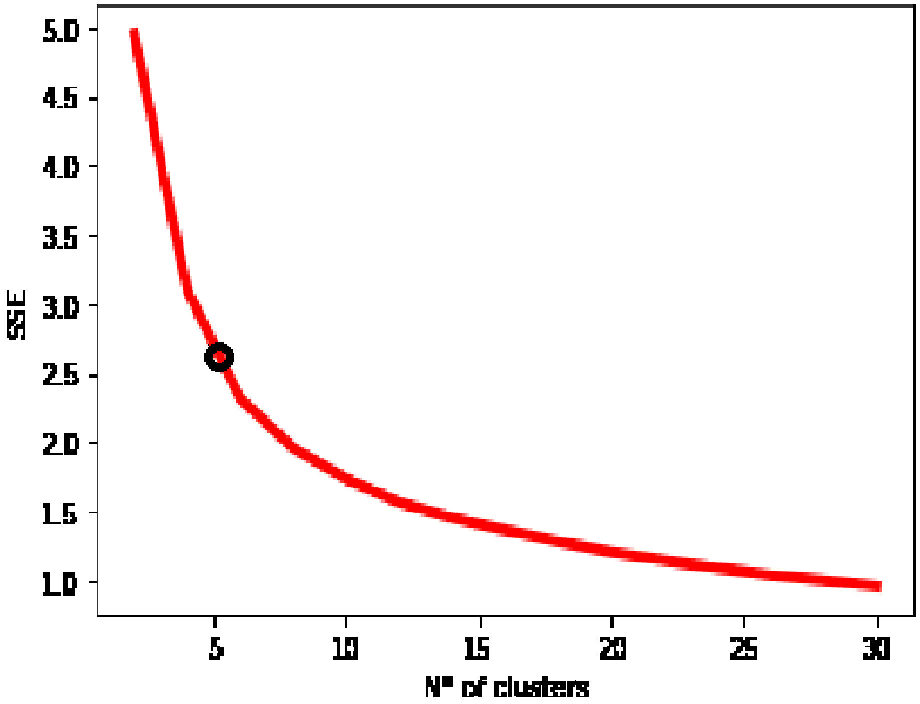


Table 8: Cluster centroids with N = 5

Cluster	PC1 +	PC2 +	PC3 -	PC4 +	PC5 -
0	-0.21	0.91	-0.71	-0.12	-0.09
1	1.68	-1.81	-1.03	0.26	-0.50
2	-1.34	-2.34	0.88	-0.35	0.74
3	3.73	1.25	1.25	-0.02	0.34
4	-2.88	0.76	0.20	0.12	-0.14

Table 9: Cluster centroids with N = 7

Cluster	PC1 +	PC2 +	PC3 -	PC4 +	PC5 -
0	-2.30	-1.30	0.92	0.63	0.14
1	-3.04	1.51	-0.46	0.28	0.28
2	3.76	1.24	1.23	-0.01	0.34
3	0.11	0.94	-0.95	0.00	0.05
4	-1.58	0.59	0.47	-0.74	-0.82
5	1.57	-1.91	-0.92	0.26	-0.49
6	-1.05	-3.29	0.76	-2.45	2.65

Table 10: Cluster centroids with N = 10

Cluster	PC1 +	PC2 +	PC3 -	PC4 +	PC5 -
0	2.13	-1.38	-1.52	0.20	-0.51
1	-1.70	0.44	0.37	-0.72	-0.84
2	4.13	1.35	0.97	0.10	0.50
3	0.32	0.89	-0.82	0.05	-0.07
4	0.64	-2.73	-0.05	0.27	-0.37
5	-1.19	-3.26	0.81	-2.65	2.88
6	-3.68	1.80	-0.07	0.20	0.02
7	2.10	0.91	2.21	-0.45	-0.28
8	-2.45	-1.24	0.95	0.68	0.18
9	-1.80	1.12	-1.12	0.10	0.55

3.3.4. Theoretical comparison of methods

The purpose of this short section is to summarize and compare the theoretical results presented in this methods sections.

What is already evident is that there is no a silver bullet: between multi-criteria analysis and unsupervised methods no approach is strongly preferable to the other, from a theoretical perpective.

Instead, when analyzing the final results, the aim of this work is to use those methods in a complementary fashion.

Altough the supervised approaches involve several domain experts, there is no guarantee that the final results are not exposed to subjective attitudes (Tanim, 2022).

Another pitfall is that, while representing the informations in a unique index comes handy for the syntetization of data, the spatial distribution and cohesistence of the risk drivers, which at the end contributed in the formation of the final metric, is not explicited. In simpler words, it could result difficult to understand why an area is classified as risky or not.

A possibility could be to consider the aggregation of main categories of variables, weighted according to the participative procedure, and check which of them has the greatest importance in the formation of a spatial risk cluster.

However, if this is the objective of the analysis, unsupervised clustering techniques appear, by definition, as more suited to achieve it. Altough these are not as useful as MCA to compute risk metrics, the fact that their results are not determined by subjective judgment, but only interpred retrospectively, should ensure an improvement in terms of objectivity in the spatial categorization of areas.

As mentioned in the previous section, some works have presented attempts to objectively derive the risk weights with “semi-unsupervised” techniques. However, those techniques still required some an active role by the analysts (Tanim et al, 2022, Dossous, 2021; Wu, 2021). Those techniques mostly employed the analysis of principal components, followed by a weighting step. This last step was conducted with different modalities in each of those works.
The portion of total variance explained by the variables was always among the main weighting criteria; while this also plays for sure an important role in the separation of clusters in K-means, giving to this statistic a determinant role in risk attribution, as would have been done building a single indicator, could have been not appropriate with the structure of the data employed in this analysis.

This idea mainly comes from the empirical observation that, given the strength in the linear correlations between socio-economic variables, which could be summarized with the first two components of pca, to some, mainly physical variables, which have been frequently judged among the important variables in supervised approach[42], would not have been given as much weight as in the supervised approach.
This fact is due to the unbalanced representation of some categories of variables in the dataset[43]

In any case, given the observation that the employment of MCA methods is consolidated in the literature (Tanim, 2022; Giupponi et al., 2015), the analysis in itself will be driven by this framework.
The main results obtained with the OWA, will be interpreted in the light of clustering results, if possible.

In the comparsion of the results. care should be taken to the fact that, as reported in section 3.3.3., three variables have been excluded from the clustering analysis, due to their evident violations of the assumptions behind the methodology employed.

There is the possibility that those variables are not well represented from the variables included in the dataset, since their exclusion does not stem from an analysis of the correlations. Also considering the fact that the analysis of correlations reduced the number of variables for MCA, the metrics employed for the production of the results are slightly different between the approaches.

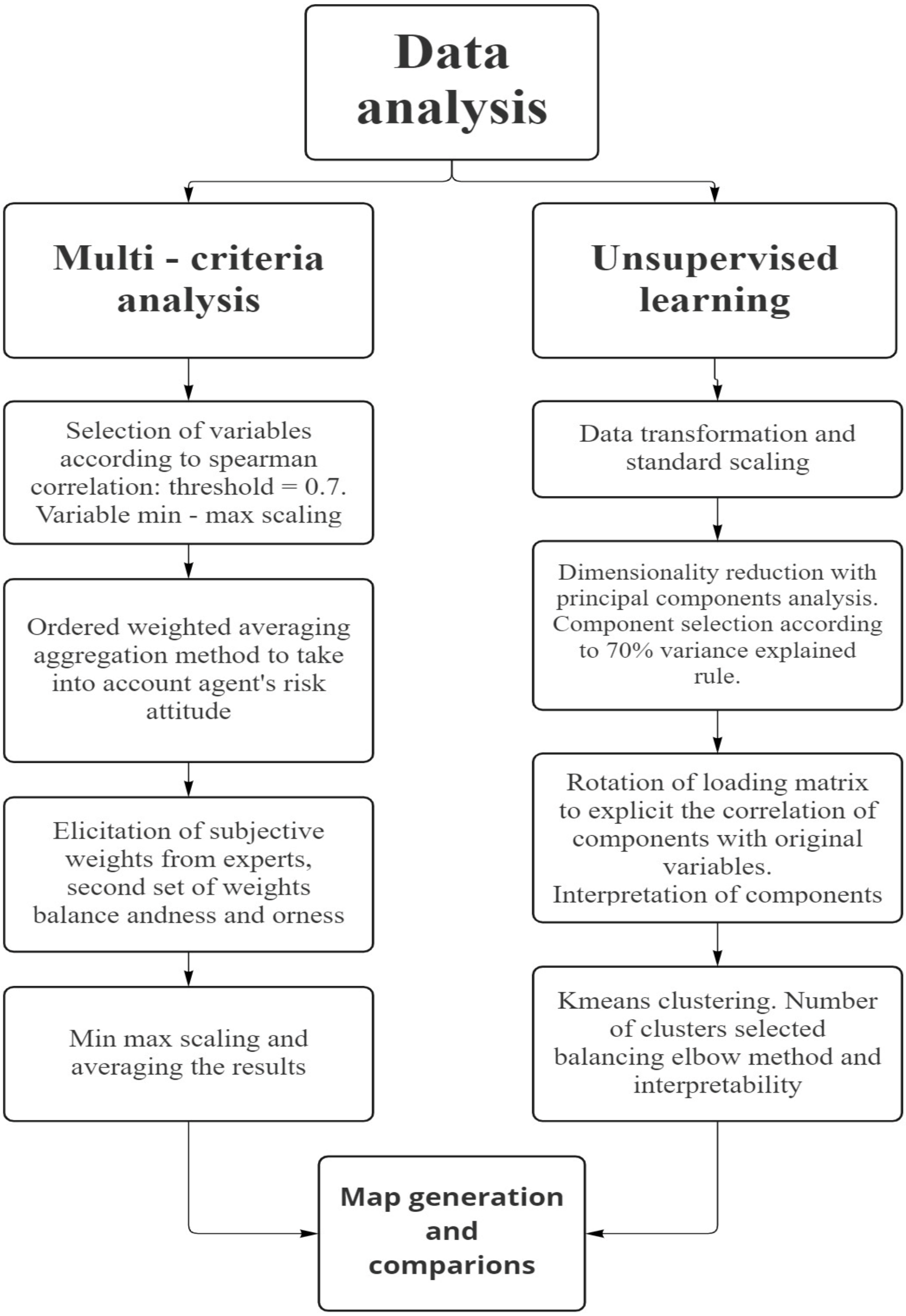
In any case I have chosen to keep the maximum possible number of variables for both approaches, also to highlight the fact that difference in the results could come from the difference in the technical requirement of the methods employed.

A comparison of the analytical methods already employed in the literature should also take care of this technical aspects, according to the theoretical results presented until now.

[42] Such as elevation, for instance

[43] This is not a problem in itself, since those variables are selected according to the literature. In fact in MCA the experts are aware of this phenomena and would adjust the weight accordingly, but in the supervised approach this problem should be considered.

Figure 12: Data analysis simplified flowchart


Data analysis
Multi - criteria analysis
Unsupervised learning
Selection of variables according to spearman correlation: threshold = 0.7. Variable min - max scaling
Data transformation and standard scaling
Ordered weighted averaging aggregation method to take into account agent's risk attitude
Dimensionality reduction with principal components analysis. Component selection according to 70% variance explained rule.
Elicitation of subjective weights from experts, second set of weights balance andness and orness
Rotation of loading matrix to explicit the correlation of components with original variables. Interpretation of components
Min max scaling and averaging the results
Kmeans clustering. Number of clusters selected balancing elbow method and interpretability
Map generation and comparions

4. Results

This section is entirely dedicated to the discussion of the results, mostly in the form of maps, obtained with the methods presented in the previous section.
The multi-criteria analysis of risk will be followed by an in-depth analysis of the risk clusters.

4.1. Multiple criteria analysis results

The presentation of the results from the Multi-Criteria Analysis framework adopted to construct a single sea level rise risk indicator begins with three global scale raster maps of risk.

The different maps represent different risk attitudes of the decision maker, which, in the OWA framework adopted, could be: 1) risk taker, 2) risk neutral and 3) risk averse.

Starting from a native resolution of 1km, those maps have been aggregated for visualization purposes at a resolution of 50 km. The aggregation function employed is the average.

Figure 13: Global risk index map, risk taker

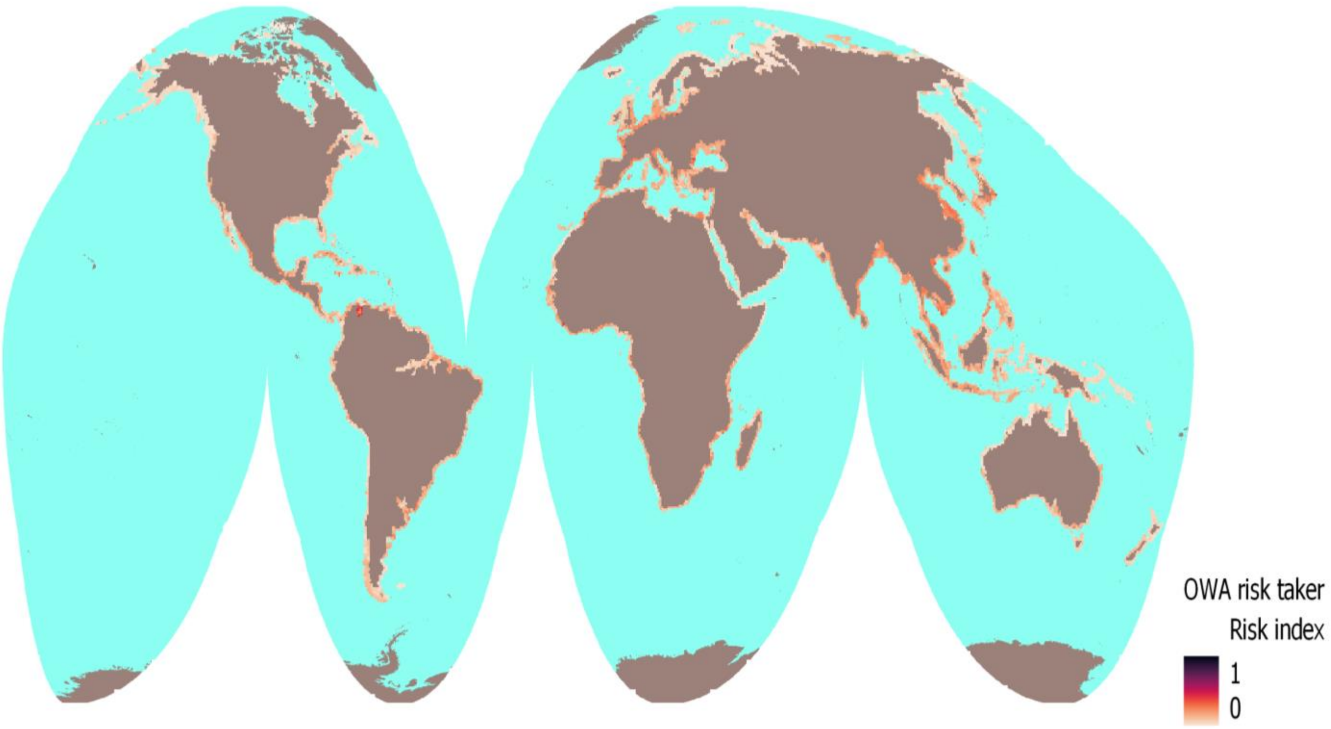


Figure 14: Global risk index map, risk neutral

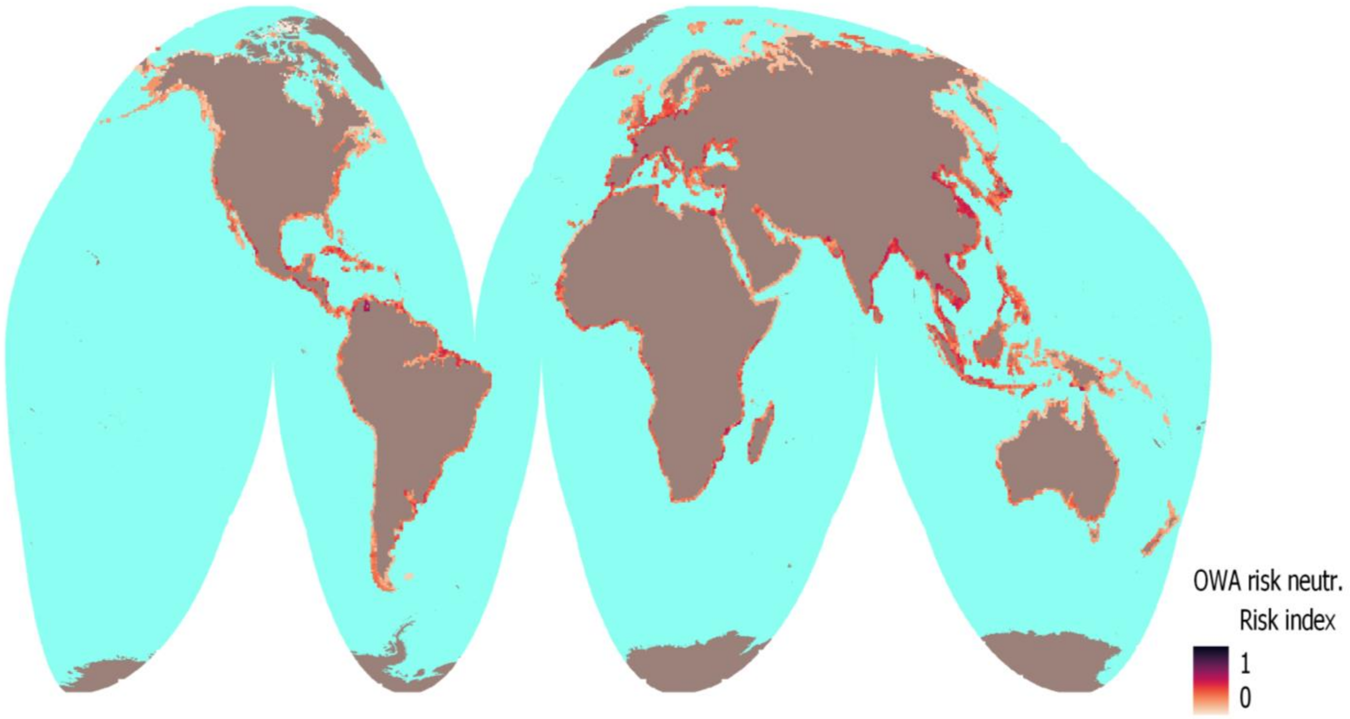
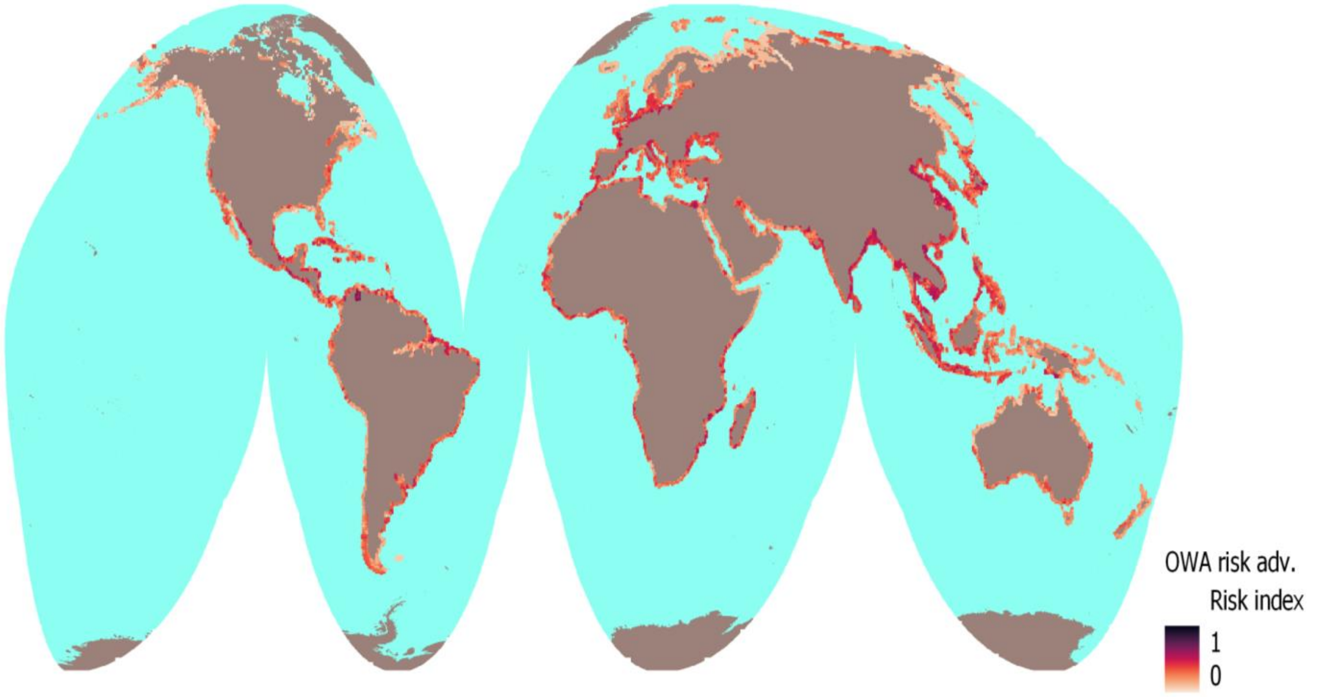


Figure 15: Global risk index map, risk averse



Some charachteristics of these maps could be highlighted, already at a global scale. Selecting the same color ramp for all the maps, ranging from 0 as mimimum to 1 as maximum theoretical values of the index, results in the visually discernible increase in the values of risk index following the increase in the degree of risk aversion.

A 0-1 min max normalization is employed twice during the MCA: in the first time to transform the raw variables and in the second time with the results obtained for each set of weights chosen by a domain expert. Nevertheless, after the final aggregation no min max transformation is applied, until now.

This was done on purpose to inspect until which extent the experts' judgements on risk agree, under different risk attitudes.
For instance, if all the experts had agreed on the attribution of the maximum value of risk to a cell, the final value for this cell should result exactly equal to 1.
This never happens because, when averaging the experts' indexes, a compensation effect between the different subjective opinions seems to takes place.
As a consequence, no cell reaches the maximum value of risk in the aggregated final index, before the normalization.

In any case, before the comparison with the risk clusters, a third min-max normalization will be applied to express the results in relative terms.

This choice resuls in similar minimum values for the three risk propensions scenarios, all indistinguishable from zero, but in different maximum values of risk, ranging from 0.75 in case of risk tolerance to 0.86 in the case of risk aversion, passing through the values of 0.8 in the case of risk neutrality.
This could be reasonable, considering the concept of risk propensity.

The results from the risk neutralitly scenarios, from a visual perpective, appear to be slightly closer to the results from the risk aversion scenario than to the results from the risk tolerance scenario. This could be due to the choice of color palette or to the accentuayed "flattening" in the distribution of the absolute risk indicator in case of risk tolerance.

South eastern Asia, the Mediterranean area, some portions of the atlantic coast of South America and the Carribean area always seem to present the relative higher values of risk, even if this is particularly evident in the case of risk aversion map. Also some parts of the African coasts could be likely identified as risk hot-spots
On the other hand, the coasts of Russia and northern America, especially getting closer to the arctic, always present low values of risk.

It is indeed possible to notice correlations in the distribution of the values in those maps: the same relative level of risk tends to be attributed to the same areas across all the maps, netting out the difference in absolute values

In any case, to better inspect the hypothesis formulated after this first graphical inspection, I decided to focus on a lower geographical scale.

Data for the three risk weighting schemas is plotted for the Mediterranean region.
In this case, the native resolution of 1km is employed.

Figure 16 Map of risk index in the Mediterranean, risk taker

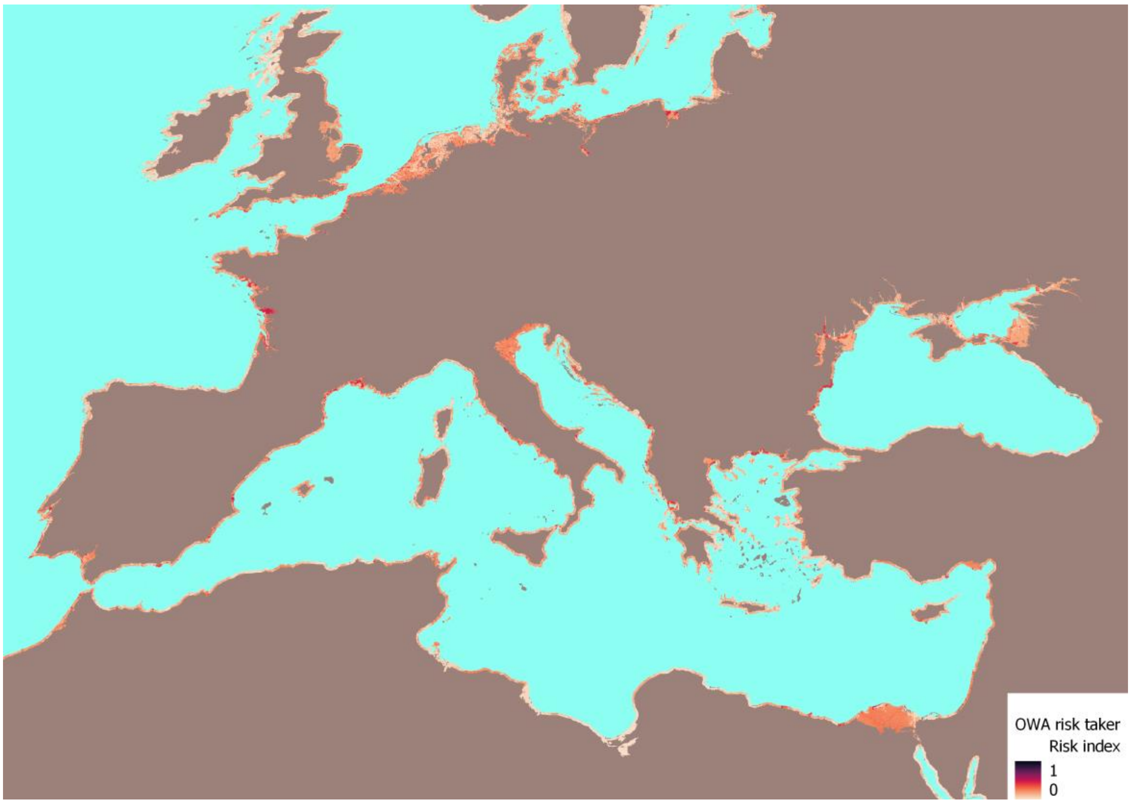


Figure 17: Map of risk index in the Mediterranean, risk neutral

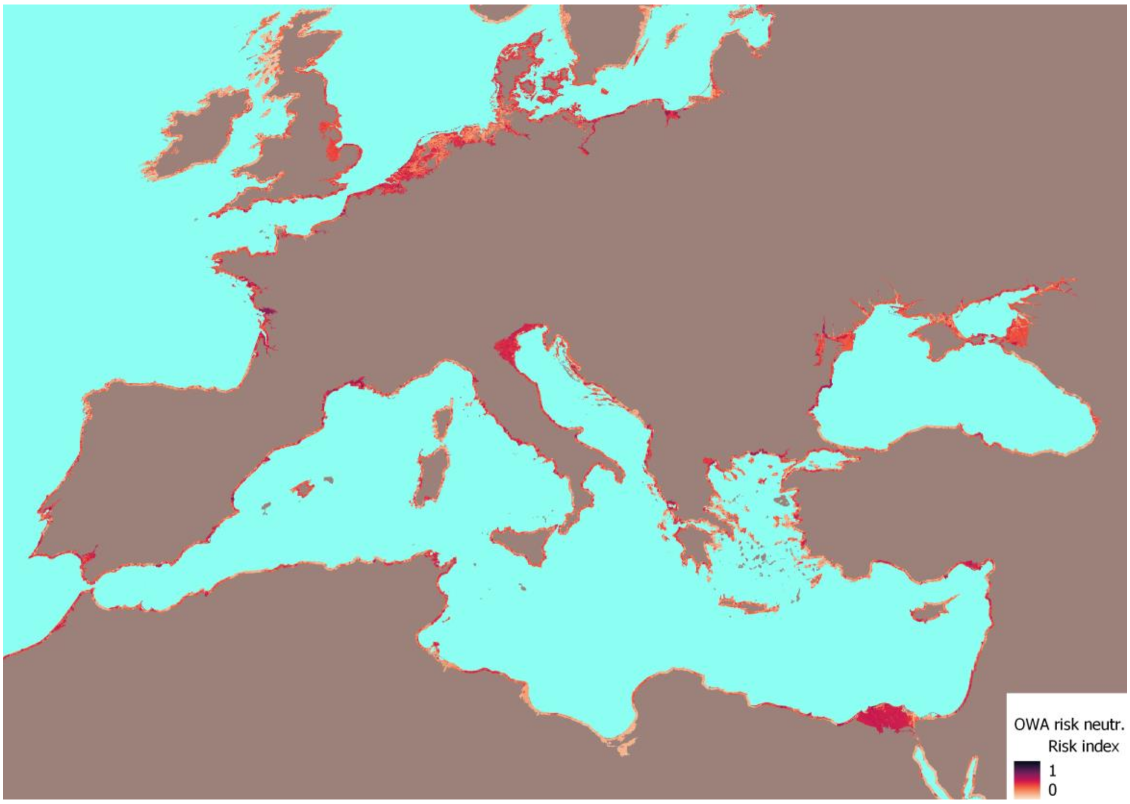
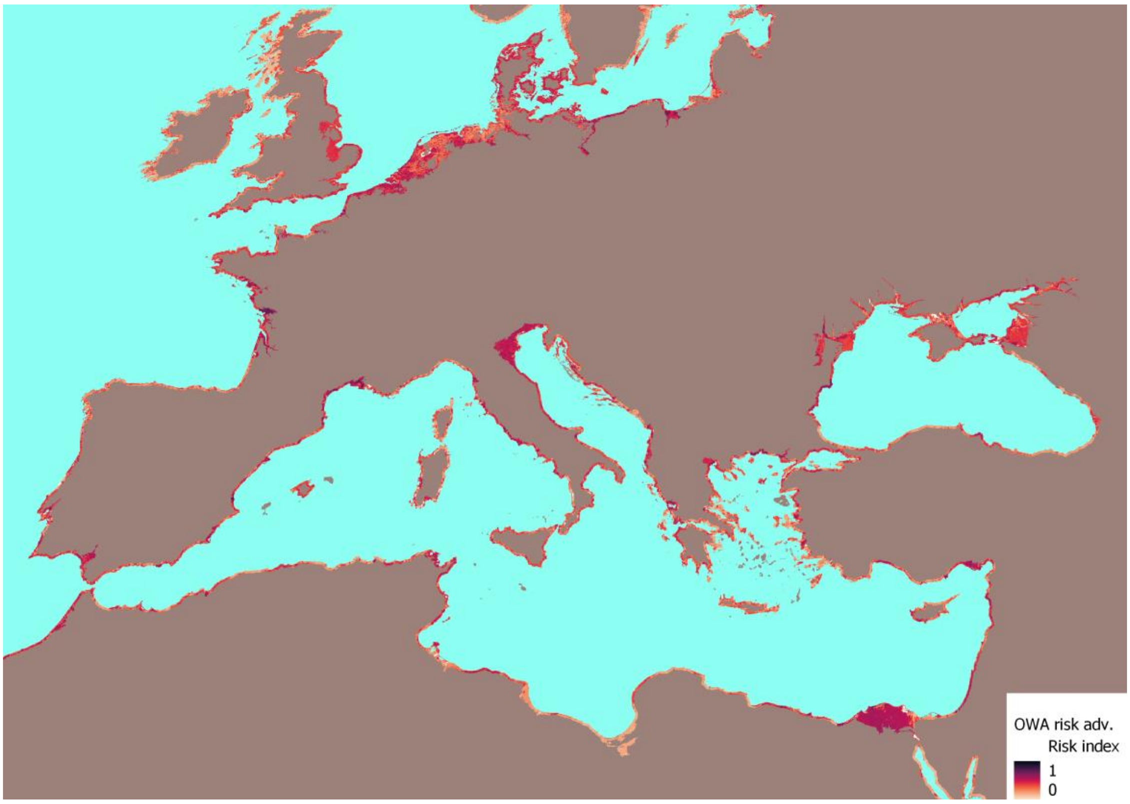


Figure 18: Map of risk index in the Mediterranean, risk averse



Examining this second set of maps, the results from the global analysis seem to be confirmed.

In detail:

- the increase in the average absolute values of the risk index is again positively associated with the degree of risk aversion

- several risk hotspots could be identified in the region, such as the great deltas of Po river in Italy and of Nile in Egypt;

- the spatial distribution of risk is similar in relative terms across the three mapping alternatives: here the correlation between the spatial distribution of the different indexes is particularly striking from a visual perspective.

Given this premises, I decided to quantify more rigorously the correlation between the values of the three different risk indicators, in the whole dataset. Spearman rank correlation coefficient was employed and the results definitely confirm the presence of a strong positive association between the indicators: none of the three couples of correlation was found lower than 0.95.

This means that, even if they can assume different absolute values, the different indicators yeld really close results in relative terms. I want to stress again the idea

that, given the (near-to) global scope of this analysis, even if the risk is defined in relative terms, the indicator would still cover almost all the possible extreme situations.
On top of that, looking at the literature, since a relative scale of risk is employed in all of the local scale assessments I cited in section 2 and 3, in this context this assumption could be even more justified.

Given the correspondence of the results between the three weightning schemas, at least in relative terms, and to simplify the workflow of the analysis, for the next maps presented I will employ only the results from the risk neutral case, being it by definition the natural midway between the other approaches

After justifying the choice of focusing on a single supervised indicator of risk, the following operation was to put an effort on an effective representation of results.
The raster map presented before are indeed difficult to employ for immediate interpretative purposes.

This operation has more than pure aesthetic scopes, the improved visualization has been used to identify regions of particular interest from the point of view of this analysis, where to focus the following comparisons between MCA and unsupervised approach results.

To produce this map, from the dataset containing the values of the risk-neutral indicator, I selected locations with values in the 0.95 quantile. In other words, the 1 km cells selected represented the top 5% of high risk locations, according to the previous assessment.

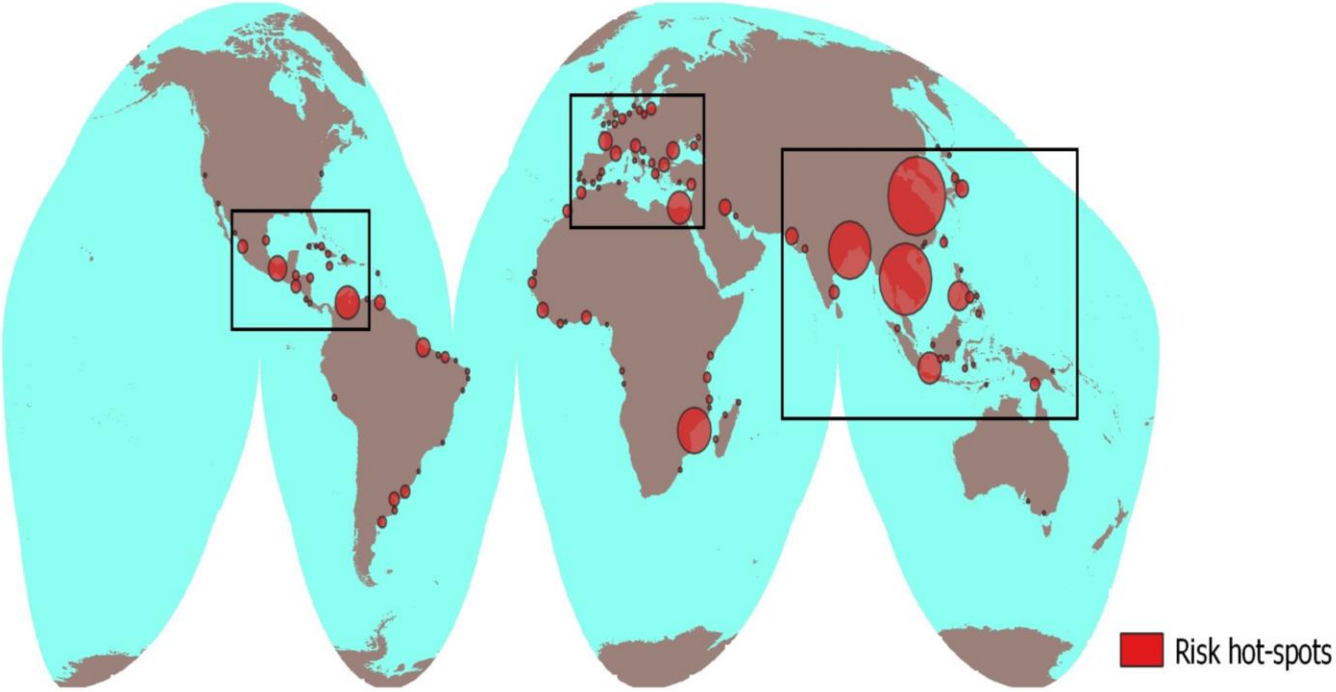
All the cell centroids which lied within 1km radius from another centroid of a high risk cell have been buffered and dissolved with Qgis native functions. This resulted in the formation of several polygons representing contiguous high risk areas.
The area of these geometrical shapes has been computed, and the polygons with an area lower than 100 km^2 have been discarded, both for reducing isolated points marked as noise and for improving the visualization of the largest contiguous high risk areas.

The polygons obtained were usually associated with irregular shapes. To addittionaly improve the representation they have been substituted with their centroids. Then, their centroids have been buffered with a radius slightly more than propotional to the original area of the figure, to accentuate the graphical representation of large high-risk areas on a global scale.

Figure 19, despite being one of the simplest maps presented here, is indeed one of the most crucial risk representations for the purposes of this thesis. Both as an important summary of what already presented and as an extremely useful starting point to determine the scope of the next analysis.

The area which will be object of more in-depth comparisons are enclosed in the rectangular shapes.

Figure 19: Global Sea level rise risk hot-spots



As already mentioned, from figure 19 it is possible to draw several conclusions.

The greatest risk hot-spots are undoubdetly located in south eastern Asia. With the three global largest high risk hot-spots located, respectively: in the coastal areas between northern China and South Korea, in Mainland Southeast Asia and in Bangladesh.

The Euro-Mediterranean region could also be collocated among the regions presenting the greatest number of hot-spots. They are mainly concentrated in the European coastal regions, even if the greatest hot-spot in the area is the delta of Nile in Egypt.

Even if the average size of threated areas is small as compared to other macro-regions, this appears as the region with the higher density of non-contiguous high-risk hotspots.

Other regions that will be object of targeted analyses are the Caraibic/Central American coastal regions. From a spatial point of view, the risk distribution in this regions appears similar to the European one. Comparing this results with the ones in figures 13-15 it is also possible to notice that in this area, and above all in the Venezuela coastlines, some of the highest values of the risk index are registered.

Another large risk hotspot is the African coastal region between South Africa and Mozambique. I have chosen to focus on this hot-spot because, if compared to the others, it does not appear inserted within a “system” of other hotspots. It was however noteworthy to mention that this appeared as one of the widest risk areas in the world.

Another particularly evident feature of this map is the almost absolute absence of coastal risk hot-spots in Northern America and in the northern part of Russia, as already found with the previuos global maps. This phenomenon is also registered in Australia and in a vast portion of the southern Atlantic coasts of Africa, where the hot-spots points are extremely rare.

As a final remainder, as already reported, the interpretation of those results should take care that this map has been graphically simplified for visualization purposes.

4.2. Unsupervised learning results

Before the comparison of the results between the different methods, some of the outcomes of unsupervised analysis are examinated individually.

To begin the discussion, as in section 4.1, I will comment the global scale map produced with three different cluster settings, selected in 3.3.3..
As for the risk index, those maps have been aggregated from 1 km to 50 km resolution for graphical purposes. In this case the aggregation function employed was the mode, being the cluster classification categorical.

Figure 20: Global clustering results, N = 5

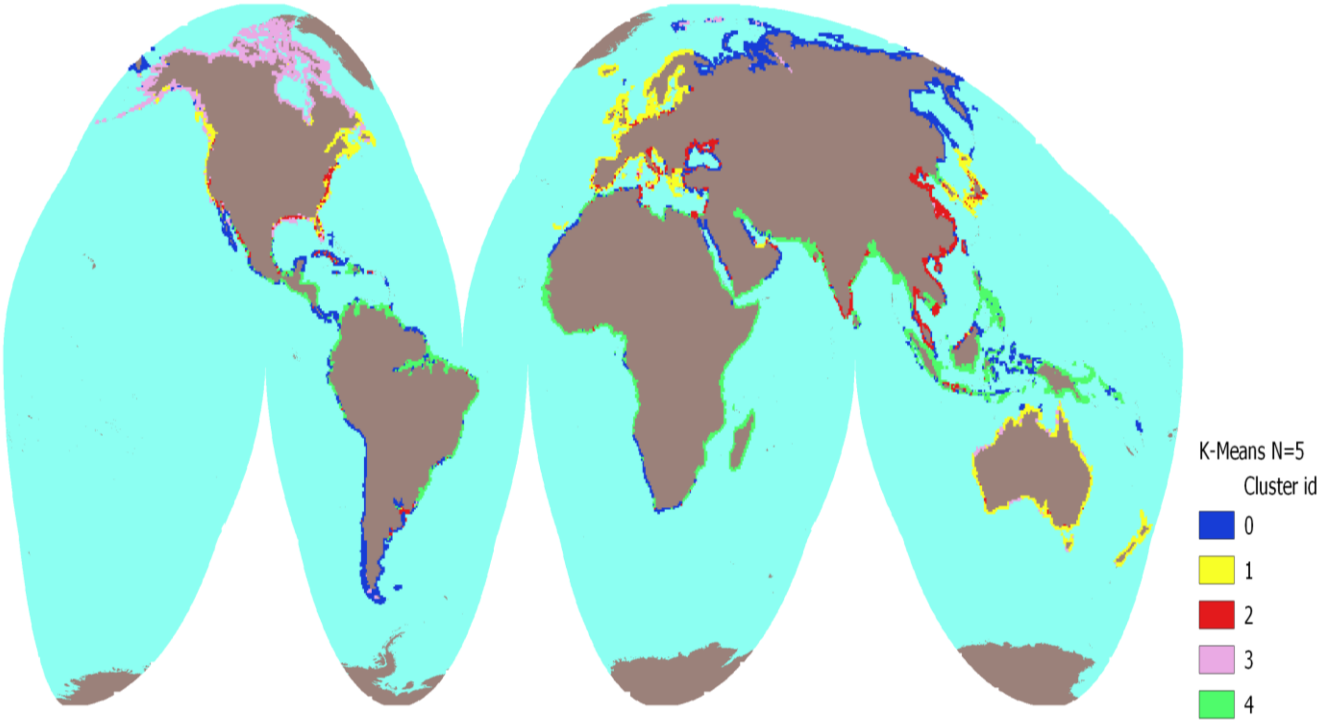


Figure 21: Global clustering results, N = 7

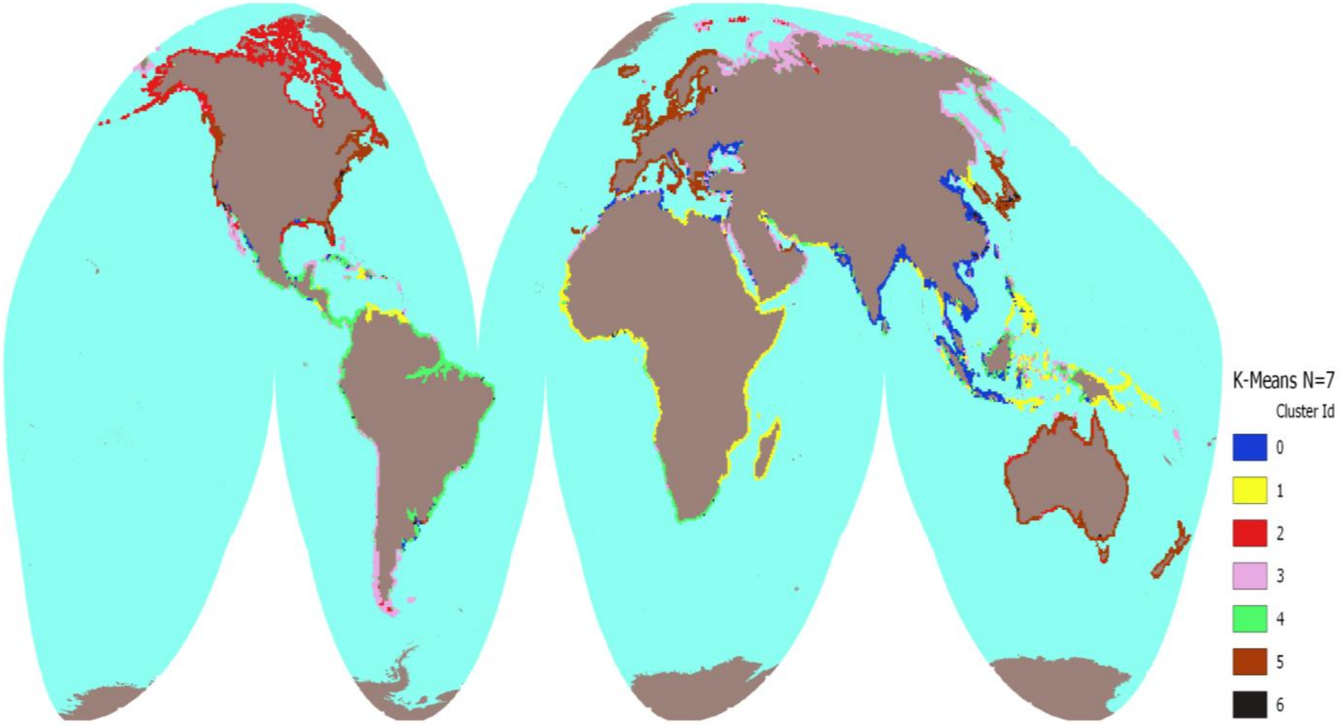
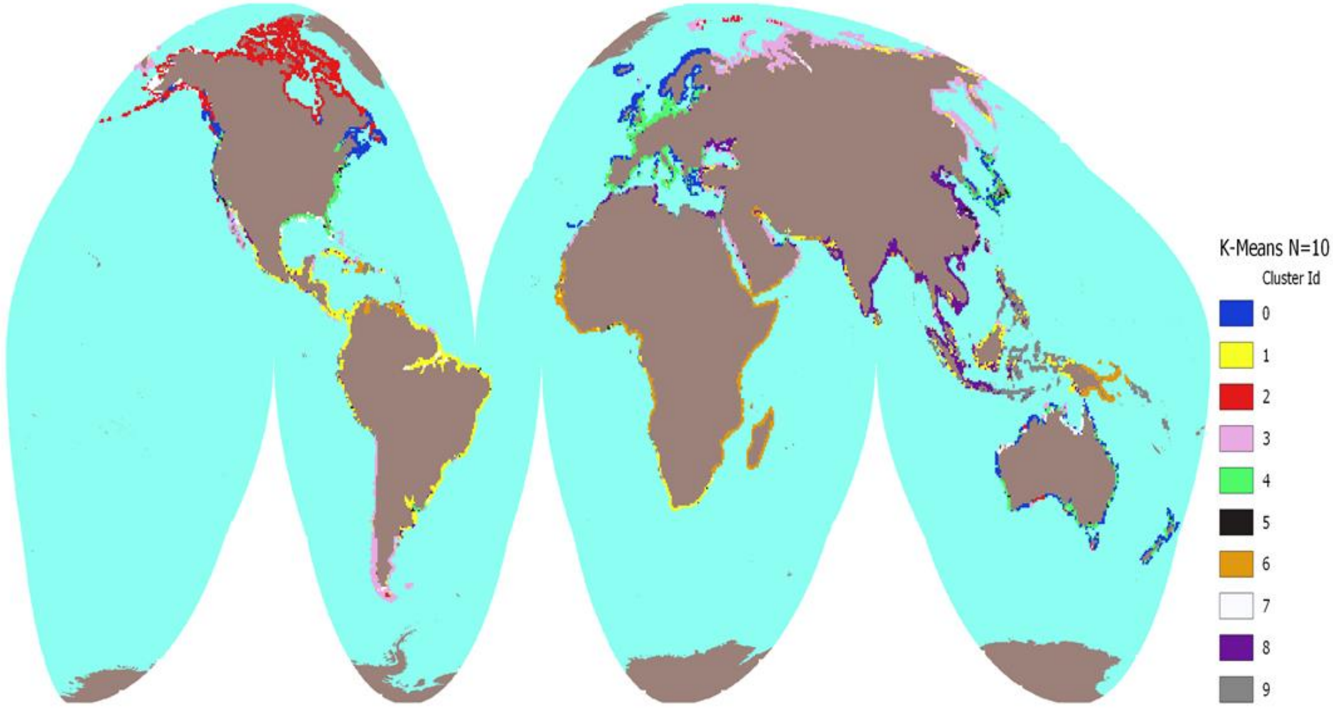


Figure 22 Global clustering results, N = 10



On a global scale, the interpretation of cluster results is even more difficult than the one for the risk index.

At a first glance, looking at figures 20-22, it is possible to distinguish the most developed areas from the others. In any case, also the contiguity in the classification in those areas is visually segmented by the presence of different smaller clusters.

Nevertheless, as explicitly stated in the conclusion of section 3, given the technical characteristics of the two different analysis, the main aim of clustering should be to add value to the interpretation the results of MCA. Mostly at the level of hot-spots.

What is already possible to notice here is that, as the number of clusters increases, the spatial separation between different groups of pixels becomes more uncertain, with different smaller clusters randomly spreaded within the same “macro-cluster”. This phenomenon will be better inspected in the Mediterranean region.

The case of Australia when the number of clusters is equal to 10 is emblematic of this issue. In this region, when the cluster number is equal to 5 or 7, almost only one cluster is identified: with N=10, the number of clusters visually explodes.

As aforementioned, to confirm these statements, a zoom of the results in the Euro-Mediterran is presented. The maps have the original resolution of 1 km.

Figure 23: Mediterranean clustering results, N = 5

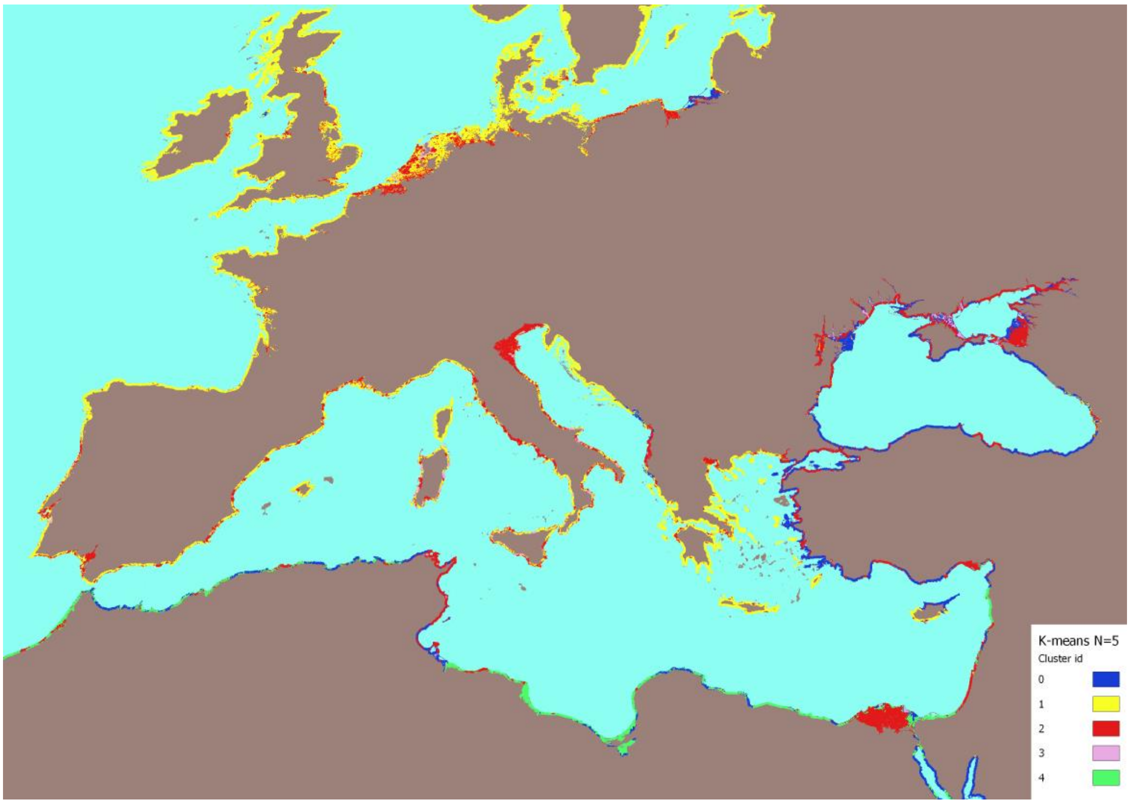


Figure 24: Mediterranean clustering results, N = 7

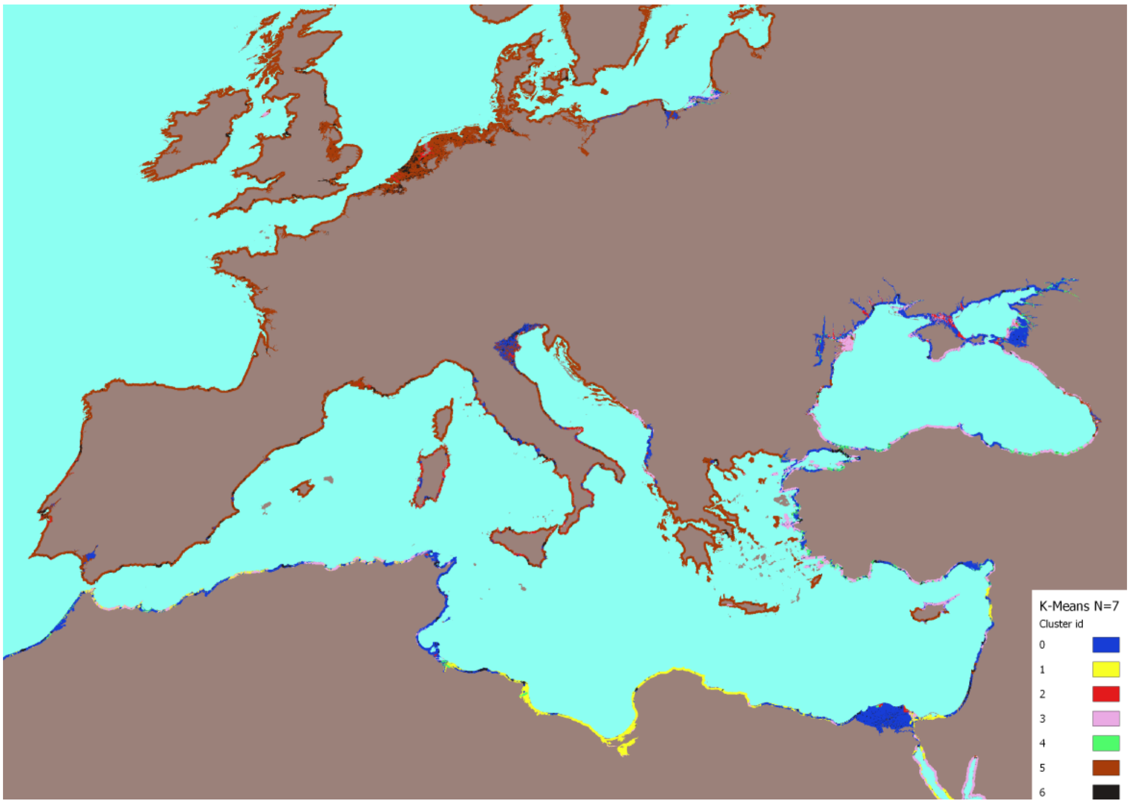
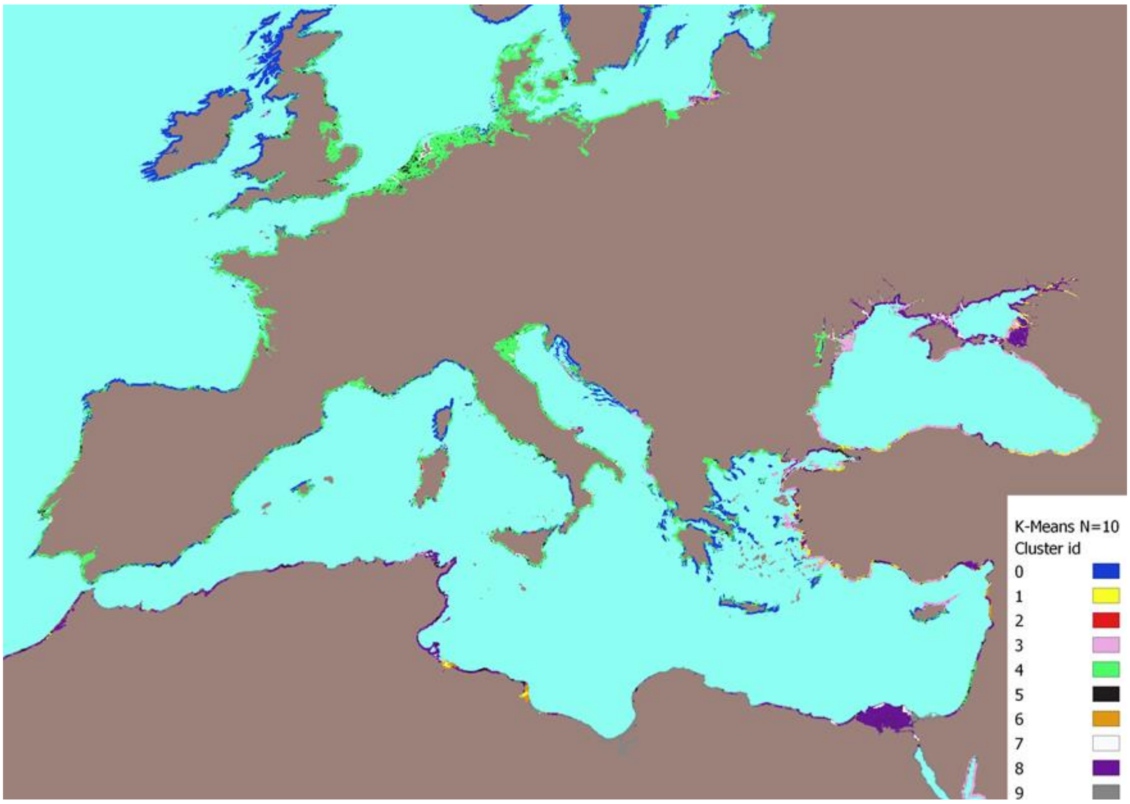


Figure 25: Mediterranean clustering results, N = 10



A closer look to Mediterranean regions confirms that difficulty in the spatial interpretation of clusters with the increase in the number of clusters.

A phenomenon which at a first glance appears to take place, mostly with N = 10, is the aggregation of clusters at the higher level of the visualization, with areas which were well separated in the case of N = 5 that now seem to form a unique macro-area. This is for instance the case of the Netherlands or of the Italian northern Adriatic coastal zones.

However, a closer focus to those areas revelas that in the N=10 case, among the visually predominant cluster it is possible to find a mosaic of smaller clusters, also within very small distances such as few square kilometers cells.

In other words, the spatial distribution of cluster at a micro-level seems characterized from a higher variability, while in the case of N = 5 the clusters are forced to be more uniform.

To conclude, what emerges from the examination of both global and local maps is that increasing the number N of clusters seems not to produce more information on the presence of difference macro-scale clusters[44].This mainly results in the introduction of additional noise between the clusters altready found when N was equal to 5.

Objectively ranking the clustering alternatives in terms of optimal spatial description of the phenomena is beyond the purpose of this thesis.
Nevertheless, considering the conclusions drawn until now in this section and the interpretative advantages given by a small number of clusters. as stated in section 3.3.3., for the comparison of unsupervised results with the MCA results I have chosen to employ the maps obtained with N = 5.

In the last part of this chapter, I want to discuss an example of the combined use of standard statistical mestric with the interpretation of spatial patterns. This approach has generally driven this analysis, and it is possible to state that it is one if its cornerstones.

As also reported in section 3.3.3., I have chosen to exclude the density of roads from the clustering of variables. This was done because even after the logarithmic transformation the variable was left severely skewed.
This could not be only found from a statistical point of view, but also from a visual perspective. In figures 26 and 27 I am presenting the comparison of the clustering results with and without roads density variable, with N = 5.

The regions selected for this comparison are the northern coastal zones of China, where the presence of main roads was clearly distinguishable.
In any case the same results could be found all over the world and for the other clustering alternatives, with N = 7 or 10. A vector layer with the roads is also plotted in dotted green lines to ease the comparison.

[44] In some case this increment achieves the outcome to reduce this information, as presented for Europe

Figure 26: Clustering without roads, China, N = 5

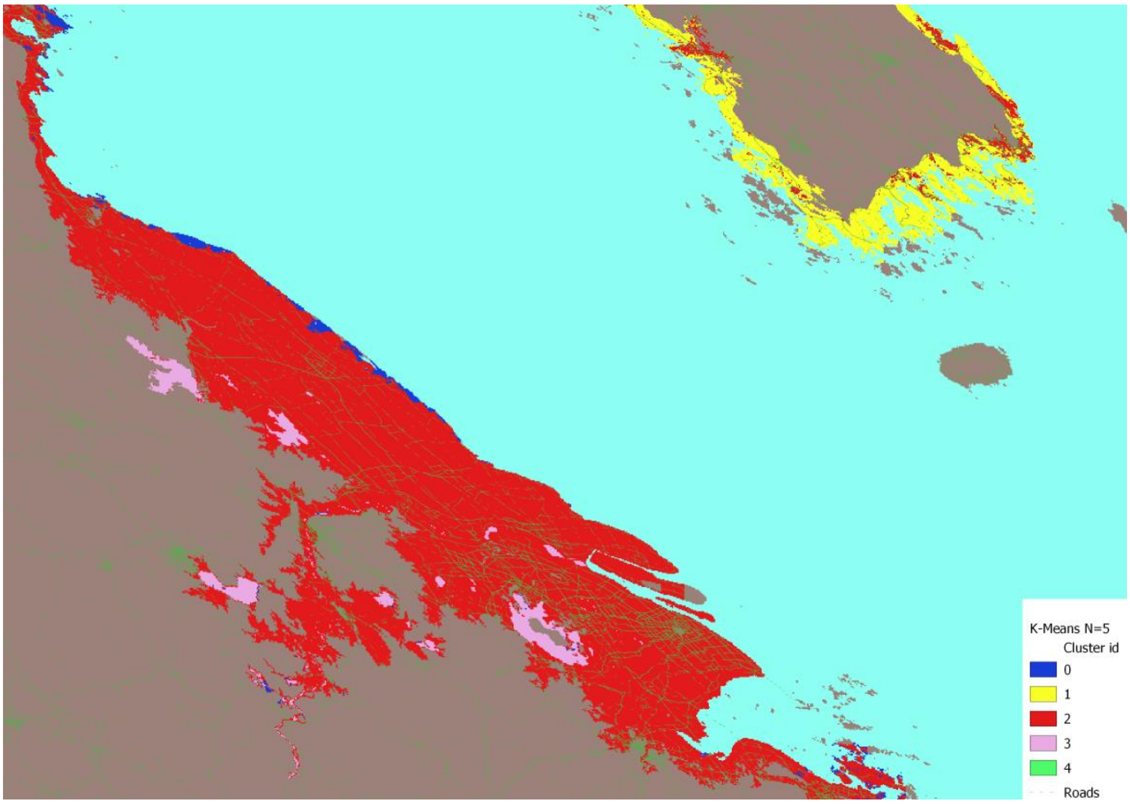
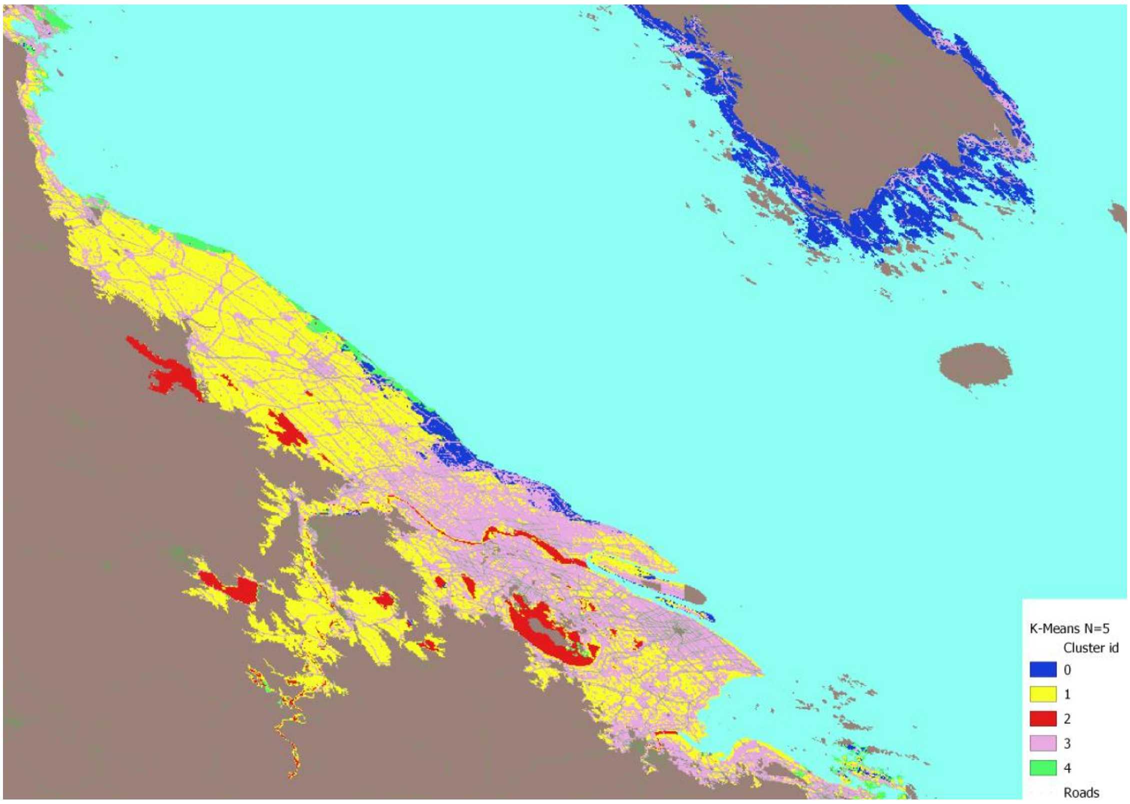


Figure 27: Clustering with roads, China, N = 5



Cluster number 3 in figure 27 (pink) is clearly driven by the presence of roads[45]
The inclusion of a single variable drammaticaly influences the results, but is uncertain if it adds more information. Since, in practice, a single cluster describes the presence of roads, it is reasonable to assume that the other variables are receiving less weight in terms of relevance in clustering aggregation.
Since the presence of urban areas are correlated with roads (see Figure 5) and this metric is already included in the supervised analysis, at this point it seemed reasonable to exclude the roads.

4.3. Comparison of results

This section will focus on the comparison of the results obtained with MCA and unsupervised learning approaches.

When comparing the results an emphasis is put on regional comparisons, with the analysis of risk in the three macro areas individuated with figure 19: South Eastern Asia, Caraibic/Central American coastal regions and Euro – Mediterranean region.

This operation has to be done because, from the global maps of risk index (fig. 16-19) and of clustering results (fig. 20-22) already presented it is rather difficult to draw general conclusions

In any case, from the global scale maps it could already be noticed that the highest values of the risk index, and consequently the greatest number of hot-spots, are registred in areas associated with the clusters labeled as 2- threatened areas or 4 – least developed areas.
Those cluster have been identified in 3.3.3. as, in order, the cluster presenting the worst risk situation in terms of average combinations of principal components and the cluster presenting extremely high values of socio-economic vulnerability.

Given the multiplicative definition of risk presented in section 3.2, the fact that threatened areas cluster, where all the risk conditions are simultaneously present, is so strongly associated with the presence of risk hot-spots, is perfectly reasonable.

Coastal areas presenting lower values of the risk index appear as splitted between different cluster and therefore it is not possible to draw similar conclusions.

I will start the analysis discussing the areas which results among the most threatened ones, according to the maps produced with all of the approaches: South Eastern Asia coastal zones.

The maps are plotted with the native resolution of 1 km.
The risk index is presented in the case of risk neutral attitude. The color scale has been modified to range from the minimum to the maximum possible value, differently from the OWA results presented before, where the maximum was kept fixed at one.

[45] The green lines with the roads are far more visually discernible in fig. 25

Figure 28: Risk index, risk neutral, South East Asia

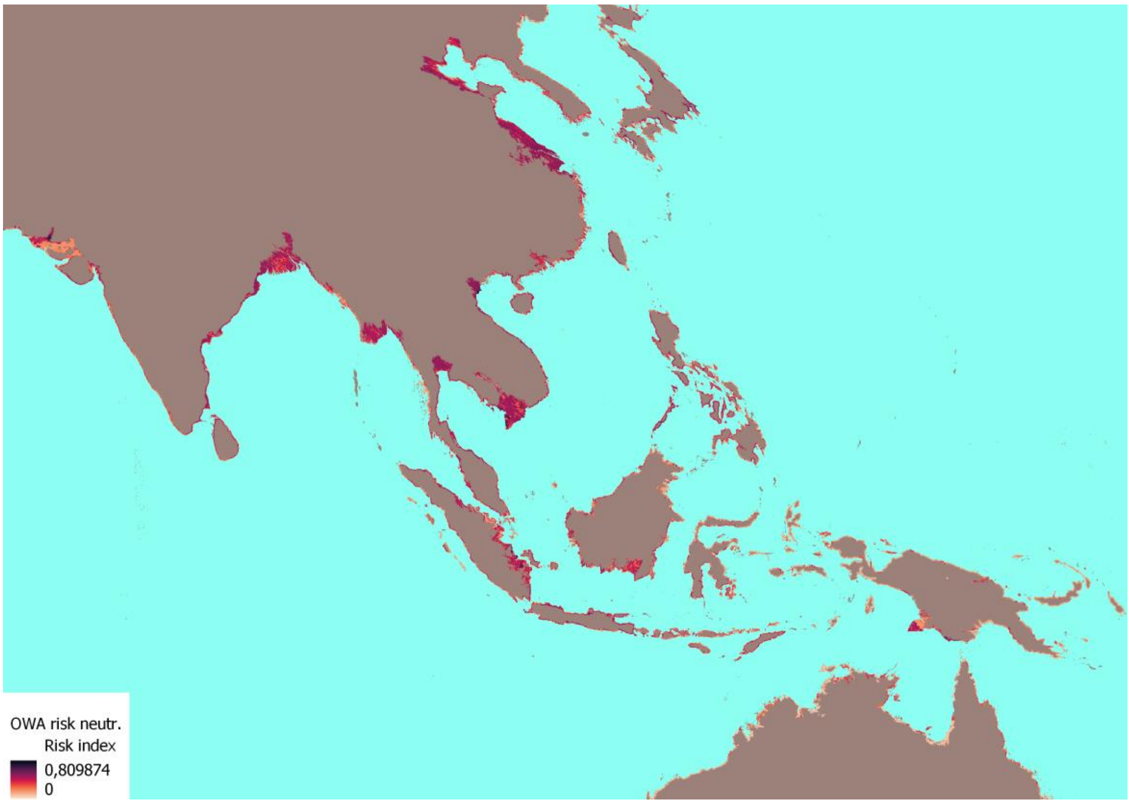
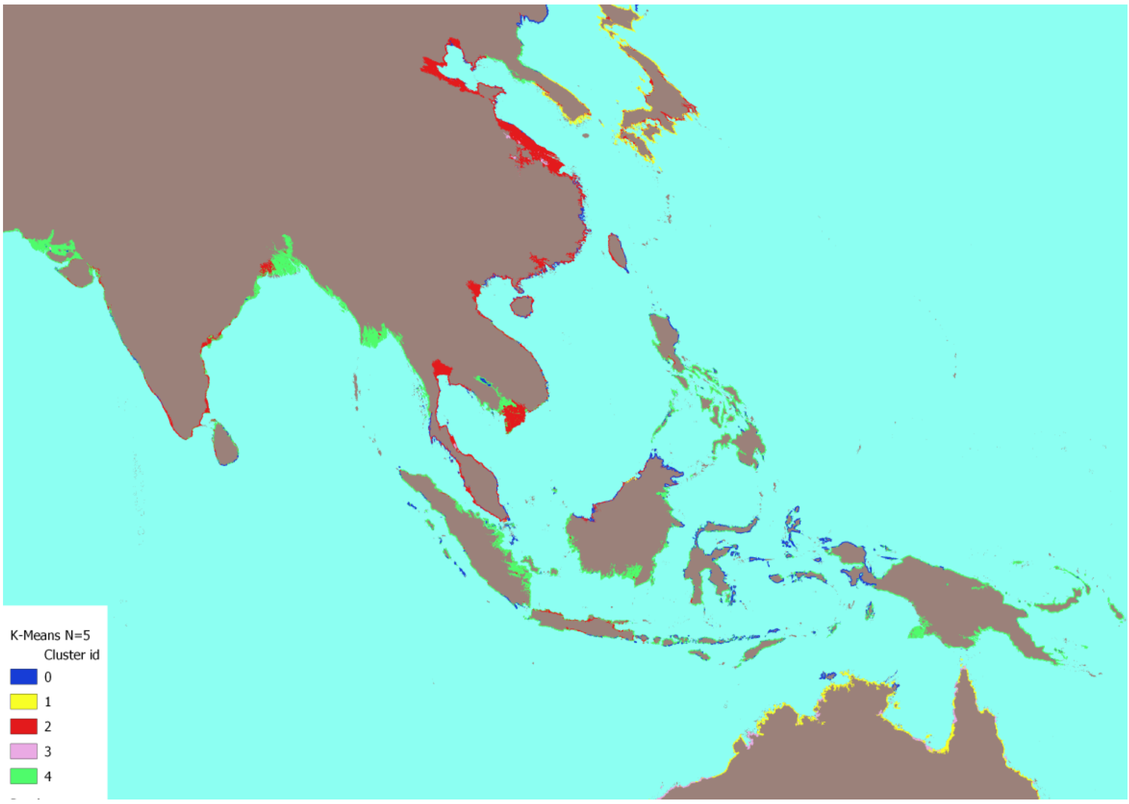


Figure 29: Clustering, N = 5, South East Asia



Two main cluster are dominant in the area, threatened areas and least developed areas cluster, which have been just discussed. This could be the reason why in this region it is possible to identify the maximum extension of risk hotspots.

In China, India and in the Indochina region the formation of risk hot-spots seems to be mainly driven by the contemporaneous presence of all the risk conditions, depicted by threatened areas cluster
In the Philippines, Malysia and Indonesia, the risk could be mainly driven by extremely poor socio-economic conditions.
In Bangladesh both clusters are located within the same risk hot-spot.
Even if cluster 0, associated with an average risk situation n according to all the components, is not well represented in this areas, several small island in the Pacific Ocean are associated with it. Those are among the areas presenting the lower risk index values in the region

Even if not properly belonging to South East Asia region, in these images it is also possible to spot some portions of Australia and Japan. This was done on purpose to highlight the contrast between those areas and the rest of the map.

In fact, differently to other areas, those are mostly assigned to cluster 1- developed areas, associated with high development and low physical vulnerability. The only criteria in that cluster pontentially contributing to an increase in risk is social exposure/vulnerability.
It is indeed possible to see that the risk values in those locations are on average low.

Another observation which emerges analyzing those maps, but could be generally true, is that while threatened areas cluster is almost always associated with high sea level rise risk, least developed areas cluster could be associated with extreme risk values, but this is not always true. The results of the other clustering options turned out to be not helpful in explaining this result.

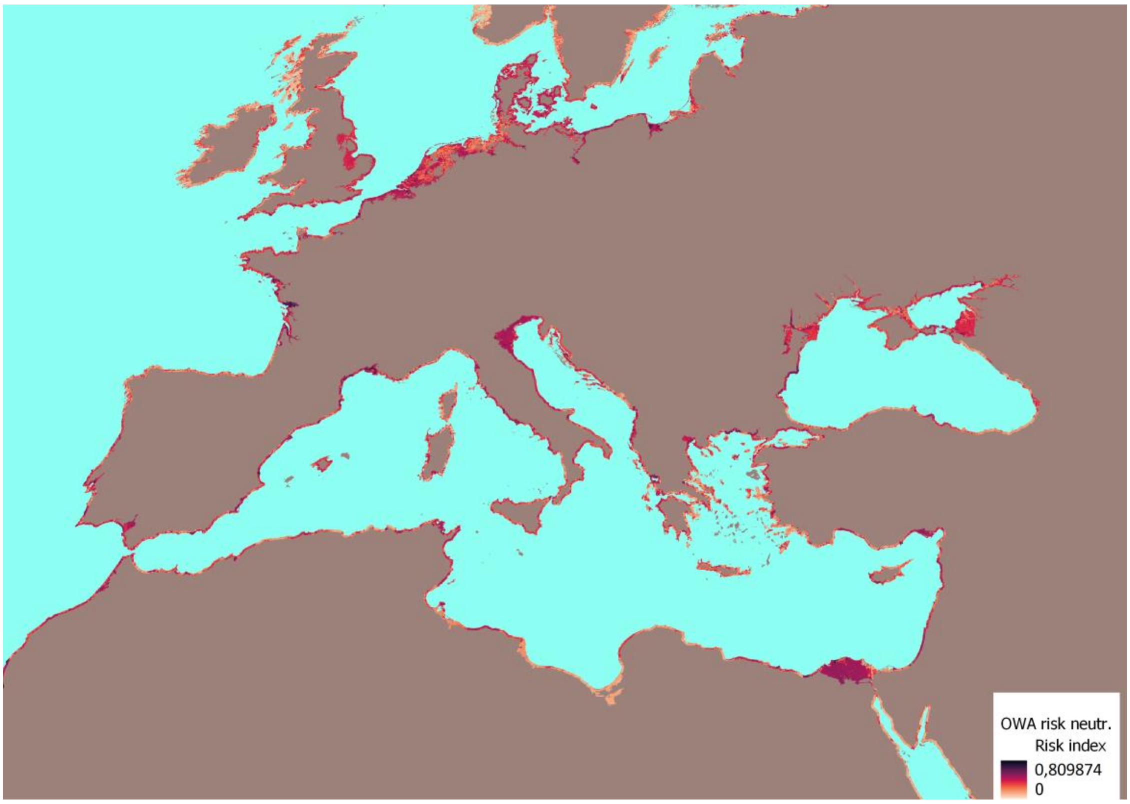
There is the possibility that the variables considered for the OWA, but not in the clustering for technical reason, play a role in this phenomena.
Another hypothesis could lie in role played by the variance of the first component, representing socio-economic criteria, in the formation of clusters with K-Means technique, as discussed in 3.3.3..
In any case, ispecting this issue in detail is beyond the scope of this thesis

The following region analyzed is the Euro-Mediterranean region. In these aresa, a high concentration of (relatively) small risk hot-spots was found; with the help, again, of Figure 19

I already plotted the clustering results for this region in Figure 23.

Even if also a map for the risk index in the Mediterranean area has been already presented (Figure 17), when comparing the different OWA settings, I decided to present another reprentation with the color palette ranging from the maximum to the minimum possible value, in Figure 30.

Figure 30: Risk index, risk neutral, Mediterranean



In the Euro-Mediterranean region it is possible to appreciate a higher variability in the clusters' spatial distribution, as compared with to the previous case-study.
To be precise, clusters 0 – average conditions, 1 – developed areas, 2 – threatened areas and 4 – least developed areas are well represented within this regions, occupying, if not equal, at least comparable proportions of the coastal areas.
Cluster 3 – optimal conditions, on the other hand is difficult to find in most of the focus areas, since it is strongly associated only with northern America

The fragmentation of socio-economic, but also physical variables, in this area could play an importan role in the peculiar distribution of risk cluster presented in figure 19. Those regions are smaller on average compared to the contiguours high risk regions in Asia[46], but are present in frequent intervals along the coastlines.

This is also confirmed from the raw values of the risk index: here, more than for other areas, high risk areas are distributed geographically closer to low risk areas, especially in north Africa.

In the case of Euro-Mediterranean area, high values of risk index appear to be mostrly associated, as in the case of South East Asia, with threatened areas cluster, while low risk values areas are mostrly associated with average conditions cluster, in less developed countries, or developed areas cluster in most developed countries.

[46] Care should also be taken on the fact that in figure 19 an emphasis is put on the graphical representation of large continuous risk areas, which gets a more than proportional weight when drawing the points-

Here, least developed areas cluster, in contrast with the previous analysises, does not seem to be associated with high risk values. In other words, socio-economic vulnerability does not appear as a driver of risk in itself within this region.

In the Netherlands, the fragmentation between high and low risk areas is striking. Looking also at the clusters, it seems that on one hand, high exposure and and physical vulnerability in threatened areas cluster and high development on the other hand have a fundamental role in this segmentation.

In the Mediterranean region more than in the others, the clustering results appear to trace the results of the risk index. The spatial fragmentation of the zones visually introduced by the different values of the risk index, is geographically similar to the sub-division of the zones obtained with the clustering.

The last region analyzed in the detail is Central-Southern America, where the distribution of hot-spots was found similar to the one in Europe, and the presence of areas with extremely high risk values has been noticed.
Also in this case the risk index in case of risk aversion and the clustering results obtained with N = 5 are reported.

Figure 31: Risk index, risk neutral, Central-Southern America

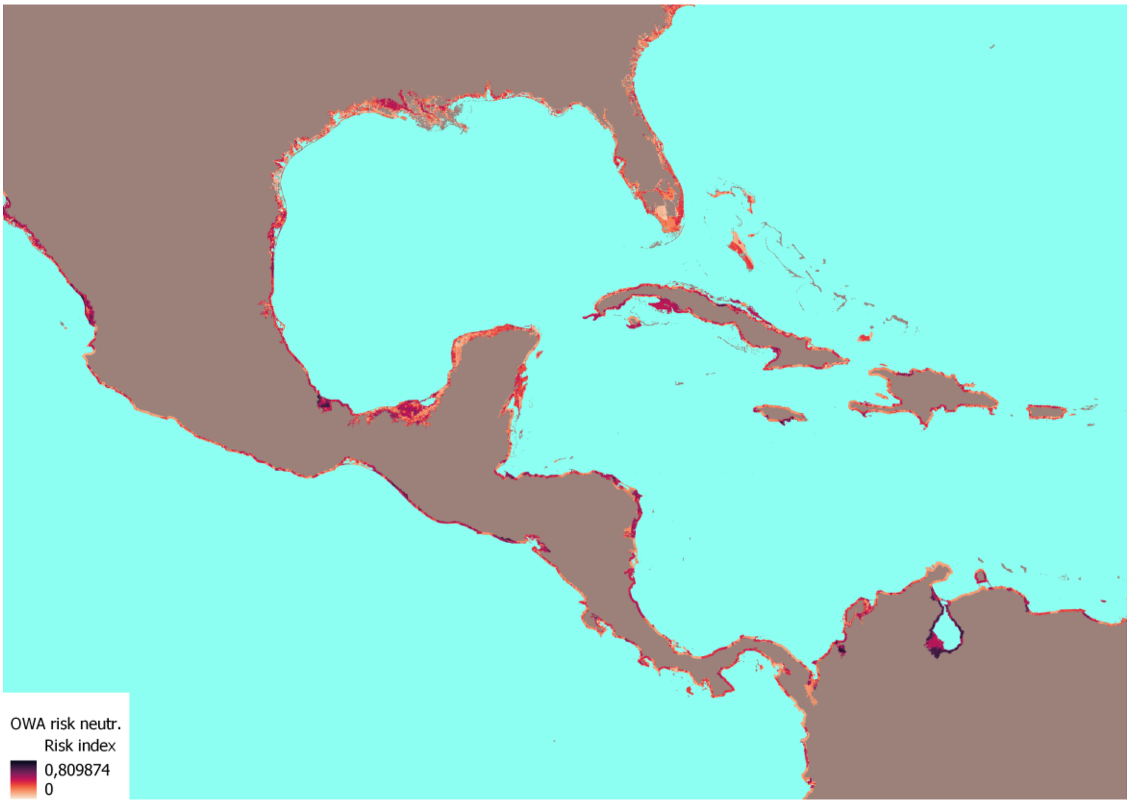
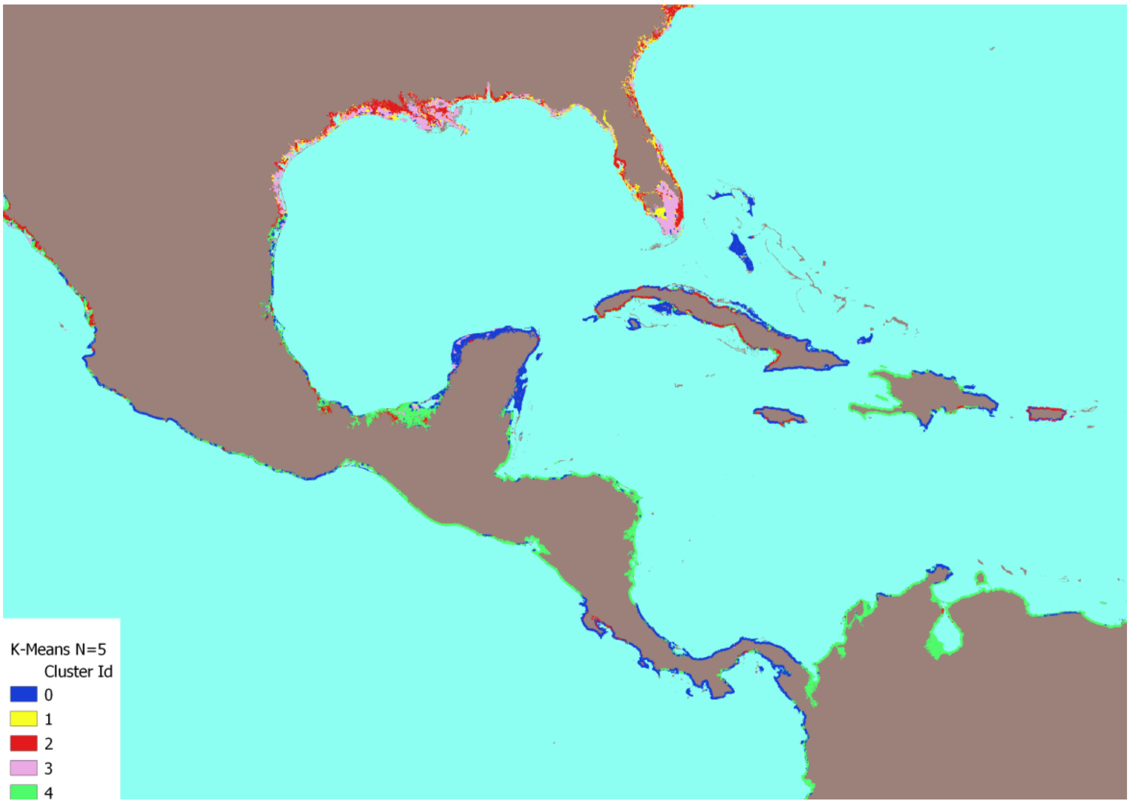


Figure 32: Clustering, N = 5, Central-Southern America



In the case of Central-Southern America, the cluster which are more represented are average conditons cluster and least developed areas cluster .

This makes the situation more comparable with the first case study, with the exception that, unlike in South-East Asia, the presence of threatened areas cluster is negligible.

Cluster number 3 make its first comparison in the northern portion of the figure, but even here is not so well represented. It is generally associated with low risk index values.

With the exception of the visible portions of the U.S. coasts, of Cuba and Mexico, where the risk appears to be mostrly driven by the presence or absence of values classified in threatened areas cluster, here the formation risk hot-spots seems to be mostly related with the extreme the socio-economic vulnerability depicted by least developed areas cluster

The case of Venezuela coastal zones is representative of this finding: it presents a concentration of very high risk areas and is, at the same time, almost entirely classified in the low socio-economics conditions cluster. Values associated with threatened areas cluster are not present within this country.

Here, both from the point of view of the risk index and of clustering, coastal zones appear more uniform, above all if compared to Europe.
In this case more than in the others, the clustering results appear as not well suited to trace the results of the risk index. This is probably due to the widespread presence of

cluster 4. The issues related with this cluster have already been discussed, and are also confirmed from this focus.

The results of this section are summarized in Table 11, which contains information on some examples of sea level rise risk hot-spots visually identified

To conclude this chapter, it is possible to state that the examination of the results substially confirmed the hypothesis formulated in the theoretical comparison of the modelling approaches

The adoption of different techniques to assess the risk has been originally driven by the idea of producing at least two different risk indicators, one subjectively determined with the help of consolidated MCA techniques and the other objectively determined with data driven approaches.

Nevertheless, an analysis of the literature and the implementation in practice of several machine learning methods revealed that, instead of being alternative to each other, those approaches should be considered complementary.

Both have their flaws. Unsupervised methods do not appear as well suited to construct single indicators while MCA results could be complex to interpret. The main conclusion remains the same as before: in this context there is no silver bullet.

Table 11: Hot – spots examples

Continent	Area	Main cluster
Asia	China / Shangdong - Hebei - Jiangsu - Lianoning; South Korea	2 - threatened areas
Asia	Cambodia - Vietnam - Thailand	2 - threatened areas
Asia	India / Western Benghal; Bangladesh	2 - threatened areas 4 - least developed areas
Asia	Indonesia / Sumatra - Java	4 - least developed areas
Asia	Philippines / Palawan	4 - least developed areas
Asia	Japan / Honshu	2 - threatened areas
Asia	Pakistan / Hyderabad; India /Gujarat	4 - least developed areas
Asia	Northern Persian Gulf (Iran; Iraq)	4 - least developed areas
Asia	India / Tamil Nadu - Andhra Pradesh	2 - threatened areas
Oceania	Southern Papua New Guinea	4 - least developed areas
Africa	North eastern South Africa; Mozambique	4 - least developed areas
Africa	Egypt / Nile delta	2 - threatened areas
Africa	Guinea Bissau; Guinea; Sierra Leon	4 - least developed areas
Africa	Ghana; Togo; Benin	4 - least developed areas
Africa	Southern Morocco	4 - least developed areas
Africa	Northern Morocco	4 - least developed areas
Europe	France / Biscay Bay	2 - threatened areas
Europe	France / Southern France	2 - threatened areas
Europe	Italy / Norther Adriatic	2 - threatened areas
Europe	Greece / Thrace	2 - threatened areas
Europe	Bulgaria; Romania	2 - threatened areas
Europe	Belgium; Netherlands	2 - threatened areas
Europe	Poland	2 - threatened areas
South America	Brazil / Amazon delta	4 - least developed areas
South America	Uruguay; Brazil / Rio Grande do Sul	2 - threatened areas
South America	Argentina / Buenos Aires; Uruguay	2 - threatened areas
South America	Argentina / Buenos Aires - Rio Negro	2 - threatened areas 4 - least developed areas
South America	Venezuela / National park of Orinoco	4 - least developed areas
South America	Western Venezuela; Northern Colombia	4 - least developed areas
South America	Southern Mexico	4 - least developed areas
South America	Cuba	2 - threatened areas

5. Summary and discussion

Several issues have been tackled in the context of this thesis, ranging from theory to interpretation. In this section I am presenting a short overview on what has been produced followed by a discussion on rooms for improvement in future analyses.

- Once identified the sea level rise hazard, it was necessary to resort to the literature, mostly to IPCC reports, to define a comprehensive risk framework and to identify the related variables.

  A multi-disciplinary approach encompassing a wide range of subjects, social sciences included, was found to be essential to carry out this work. This is reasonable given the broad range of climate changes impacts.

  Within this step is was also necessary to reason on the definition of coastal zones to be adopted.

- Incorporating a spatial dimension in the analysis has been necessary, given the above mentioned multi-disciplinariety of this research. While some phenomena could be represented at a wider scale others show a non-negligible variability within very small dimensions.

  What determines risk are the patterns in the combination of global and local scale variables.

  Hence, neither an analysis focusing on single variables, nor an aggregate scale analysis are suitable to well represent risk.

- The employment of geographical information systems (GIS) tools was needed to organize the data for further analyses. This process included a careful data selection step: thanks to the availability of several official sources it was possible to obtain informations with an accurate geographical resolution for most of the variables

  Then, to standardize those data, a raster format, a coordinate reference systems (CRS) and reprojection algorithms have been chosen.

  This was done trying to limit the unavoidable distorsions in the spatial representation of data, measured by a customized error metric.

- Multi-criteria analysis, employing a participative and expertise based approach, has been used to produce a risk-index. Unsupervised clustering techniques have been used to inspect the spatial distribution of slr - related phenomena and, consequently, to improve the interpretability of the risk index.

  Before their employement, both those techniques have been tuned on the data in order to ensure the robustness as well as the interpretability of their results.

- In practice, it was possible to identify risk hot-spots and macro-regions where the concentration of these risk hot-spots was particularly high. South eastern Asia, Euro – Mediterranean region and Central – Southern America have been inspected in detail, comparing clustering and risk index results.

Most of the risk hot-spots were identifiable either in areas with on average negative combinations in terms of risk drivers or in extremely low development territories

Given this summary, the features which could mostly contribute to the strength and novelty of this work are: its global scale, the use of a multi-disciplinary approach, the integration of the informations in a spatial framework, the comparison of different modelling approaches and an emphasis on the interpretability of the results

Obviously, there are several directions along which this analysis, which is thought as a starting point, could be improved.
Now, I am exposing them in detail:

- First of all, this is a current level analysis. Several works are now presenting results for different future scenarios, also following the indication in the latest IPCC assessment report (IPCC, 2022).
  Here, the choice of limiting the time-span to the present was mainly due to limited availability of data on future scenarios for all of the employed variables.
  Since one of the main ideas of this work was to test unsupervised methods, which are also designed to reduce data dimenstionality, in order for this kind of analysis to be meaningful it was necessary to use several metrics. This resulted in an unavoidable trade-off with the scenario analysis.

- Another related issue is the fact that sea level rise here is assumed constant, while, according to IPCC projections (IPCC, 2019) its magnitude will be differential between the global coastal zones. However, sea level rise projections are related to the different scenarios, and it is not reasonable to employ those projections without considering the consequent variation in the value of the principal, mostly socio-economic, variables.

- For some essential and directly quantificable variables it was not possible to find global data. Those variables are mostly related to active adaptation actions such as the presence of coastal protection infrastructure or the availability of funds to constrast sea level rise in threatened locations. It is evident that a complete analysis of risk should also consider adaptation practises already put in place.

- Other variables used, mostly socio-economic indicators, were available only at a national scale. I am aware of attempts to map some of those variables at a finer scale, but at the time of the analysis no related data was available.

- Dimensionality reduction techniques alternative to Pca could have been explored to include the binary variables also in the unsupervised approach

- Different and more robust clustering techniques could have been employed in the context of the supervised approach, but computational limitations at the time of the analysis did not allow an efficient application of those methods to the whole dataset.

- Given its fine resolution, the dataset could have been explored even more in depth then it was done in section 4. Analytically discussing all the risk hot-spots in detail was clearly beyond the scope of this dissertation, but I am confident in the possibility of uncovering other interesting patterns with future employements of the information contained in this data source.

6. Conclusions

This thesis represents an attempt to map global coastal zones risk in relation to sea level rise risk

The most important outcome was the attempt to establish a framework, both from a theoretical as well as from a computational point of view, to explore sea level rise risk.

With this result it was already possible to identify several threatened areas, even if this analysis remains opened to further refinements.

The core structure of this framework embedds the potentiality to be employed with the inclusion of other criterias and the variation in risk parameters.

Everything considered, hopefully this contribution has been and will be helpful in the definition and in the selection of risk hot-spots where to apply different, tailored, modelling approaches.

References

A

Addy, J. W. G., Ellis, R. H., Macdonald, A. J., Semenov, M. A., & Mead, A. (2021). Changes in agricultural climate in South-Eastern England from 1892 to 2016 and differences in cereal and permanent grassland yield. *Agricultural and Forest Meteorology, 308-309*, 108560. doi:https://doi.org/10.1016/j.agrformet.2021.108560

Alaniz, A. J., Smith-Ramírez, C., Rendón-Funes, A., Hidalgo-Corrotea, C., Carvajal, M. A., Vergara, P. M., & Fuentes, N. (2022). Multiscale spatial analysis of headwater vulnerability in South-Central Chile reveals a high threat due to deforestation and climate change. *Science of The Total Environment, 849*, 157930. doi:https://doi.org/10.1016/j.scitotenv.2022.157930

Amatulli, G., Domisch, S., Tuanmu, M.-N., Parmentier, B., Ranipeta, A., Malczyk, J., & Jetz, W. (2018). A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. *Scientific Data, 5*(1), 180040. doi:10.1038/sdata.2018.40

Arthur, D., & Vassilvitskii, S. (2007). *k-means++: the advantages of careful seeding.* Paper presented at the SODA '07.

B

Bartlett, M. S. (1951). The effect of standardization on a χ2 approximation in factor analysis. *Biometrika, 38*(3-4), 337-344. doi:10.1093/biomet/38.3-4.337

Bucherie, A., Hultquist, C., Adamo, S., Neely, C., Ayala, F., Bazo, J., & Kruczkiewicz, A. (2022). A comparison of social vulnerability indices specific to flooding in Ecuador: principal component analysis (PCA) and expert knowledge. *International Journal of Disaster Risk Reduction, 73*, 102897. doi:https://doi.org/10.1016/j.ijdrr.2022.102897

C

Cian, F., Giupponi, C., & Marconcini, M. (2021). Integration of earth observation and census data for mapping a multi-temporal flood vulnerability index: a case study on Northeast Italy. *Natural Hazards, 106*(3), 2163-2184. doi:10.1007/s11069-021-04535-w

Clemente, M. F., D'Ambrosio, V., & Focareta, M. (2022). The proposal of the Coast-RiskBySea: COASTal zones RISK assessment for Built environment bY extreme SEA level, based on the new Copernicus Coastal Zones data. *International Journal of Disaster Risk Reduction, 75*. doi:10.1016/j.ijdrr.2022.102947

D

Dossou, J. F., Li, X. X., Sadek, M., Sidi Almouctar, M. A., & Mostafa, E. (2021). Hybrid model for ecological vulnerability assessment in Benin. *Sci Rep, 11*(1), 2449. doi:10.1038/s41598-021-81742-2

E

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.* Paper presented at the KDD.

F

Finch, A. P., Brazier, J. E., Mukuria, C., & Bjorner, J. B. (2017). An Exploratory Study on Using Principal-Component Analysis and Confirmatory Factor Analysis to Identify Bolt-On Dimensions: The EQ-5D Case Study. *Value in Health, 20*(10), 1362-1375. doi:https://doi.org/10.1016/j.jval.2017.06.002

Frank, A. F. (2010). The Gini Index and Measures of Inequality. *The American Mathematical Monthly, 117*(10), 851-864. doi:10.4169/000298910x523344

Füssel, H.-M. (2007). Vulnerability: A generally applicable conceptual framework for climate change research. *Global Environmental Change, 17*(2), 155-167. doi:10.1016/j.gloenvcha.2006.05.002

G

Ghahramani, Z. (2004). Unsupervised Learning. In O. Bousquet, U. von Luxburg, & G. Rätsch (Eds.), *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures* (pp. 72-112). Berlin, Heidelberg: Springer Berlin Heidelberg.

Giupponi, C., Mojtahed, V., Gain, A. K., Biscaro, C., & Balbi, S. (2015). Chapter 6 - Integrated Risk Assessment of Water-Related Disasters. In J. F. Shroder, P. Paron, & G. D. Baldassarre (Eds.), *Hydro-Meteorological Hazards, Risks and Disasters* (pp. 163-200). Boston: Elsevier.

Gonzalez, C. A. D., Calderon, Y. M. M., Cruz, N. A. M., & Sandoval, L. E. P. (2022). Typologies of Colombian off-grid localities using PCA and clustering analysis for a better understanding of their situation to meet SDG-7. *Cleaner Energy Systems*, 100023. doi:https://doi.org/10.1016/j.cles.2022.100023

Goodchild, M. F., & Kemp, K. K. (1990). *NCGIA core curriculum*. Santa Barbara, Calif.: National Center for Geographic Information and Analysis.

Gornitz, V. (1991). Global coastal hazards from future sea level rise. *Palaeogeogr. Palaeoclimatol. Palaeoecol., 89,* 379-398. doi:10.1016/0031-0182(91)90173-O

H

Haasnoot, M., Kwakkel, J. H., Walker, W. E., & ter Maat, J. (2013). Dynamic adaptive policy pathways: A method for crafting robust decisions for a deeply uncertain world. *Global Environmental Change, 23*(2), 485-498. doi:https://doi.org/10.1016/j.gloenvcha.2012.12.006

Hossain, S. K. A., Mondal, I., Thakur, S., & Fadhil Al-Quraishi, A. M. (2022). Coastal vulnerability assessment of India's Purba Medinipur-Balasore coastal stretch: A comparative study using empirical models. *International Journal of Disaster Risk Reduction, 77,* 103065. doi:https://doi.org/10.1016/j.ijdrr.2022.103065

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24,* 417-441. doi:10.1037/h0071325

I

IPCC. (2012). *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change*: Cambridge University Press.

IPCC. (2013). Climate change 2013: The physical science basis, in contribution of Working Group I (WGI) to the Fifth Assessment Report (AR5) of the Intergovernmental Panel on Climate Change (IPCC).

IPCC. (2014). AR5 WGII Technical Summary. In (pp. 35-94).

IPCC. (2019). Integrative Cross-Chapter Box on Low-lying Islands and Coasts. In *The Ocean and Cryosphere in a Changing Climate* (pp. 657-674).

IPCC. (2021). AR6 WGI Technical Summary. In (pp. 33-144).

IPCC. (2022). Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. In *Climate Change 2022: Impacts, Adaptation, and Vulnerability.*: Cambridge University Press. In Press.

IPCC, & Houghton, J. T. (1990). *IPCC first assessment report*. Geneva: WMO.

J

Janssen, V. (2009). Understanding coordinate reference systems, datums and transformations. *International Journal of Geoinformatics, 5*.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202. doi:doi:10.1098/rsta.2015.0202

K

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35*(4), 401-415. doi:10.1007/BF02291817

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark Iv. *Educational and Psychological Measurement, 34*(1), 111-117. doi:10.1177/001316447403400115

Ketchen, D. J., & Shook, C. L. (1996). The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique. *Strategic Management Journal, 17*(6), 441-458. Retrieved from http://www.jstor.org/stable/2486927

Kokoska, S., & Zwillinger, D. (2000). CRC Standard Probability and Statistics Tables and Formulae, Student Edition. CRC Press. https://doi.org/10.1201/b16923

L

Lavalle, C. (2011). Coastal Zones In C. Rocha Gomes, C. Baranzelli, & F. Batista e Silva (Eds.), *Policy alternatives impacts on european coastal zones 2000-2050.*: Publications Office of the European Union.

Le Cozannet, G., Garcin, M., Bulteau, T., Mirgon, C., Yates, M. L., Méndez, M., . . . Oliveros, C. (2013). An AHP-derived method for mapping the physical vulnerability of coastal areas at regional scales. *Nat. Hazards Earth Syst. Sci., 13*(5), 1209-1227. doi:10.5194/nhess-13-1209-2013

Lee, S., Kim, J.-C., Jung, H.-S., Lee, M. J., & Lee, S. (2017). Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. *Geomatics, Natural Hazards and Risk, 8*(2), 1185-1203. doi:10.1080/19475705.2017.1308971

Lempert, R. J., Popper, S. W., & Bankes, S. C. (2003). *Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis*. Santa Monica, CA: RAND Corporation.

M

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations.*

Magnan, A., Schipper, L., Burkett, M., Bharwani, S., Burton, I., Eriksen, S., . . . Ziervogel, G. (2016). Addressing the risk of maladaptation to climate change. *Wiley interdisciplinary reviews: Climate Change, 7*, 646-665. doi:10.1002/wcc.409

McGranahan, G., Balk, D., & Anderson, B. (2007). The rising tide: assessing the risks of climate change and human settlements in low elevation coastal zones. *Environment and Urbanization, 19*(1), 17-37. doi:10.1177/0956247807076960

Meijer, J., Huijbregts, M., Schotten, K., & Schipper, A. (2018). Global patterns of current and future road infrastructure. *Environmental Research Letters, 13*. doi:10.1088/1748-9326/aabd42

Merkens, J.-L., Reimann, L., Hinkel, J., & Vafeidis, A. T. (2016). Gridded population projections for the coastal zone under the Shared Socioeconomic Pathways. *Global and Planetary Change, 145*, 57-66. doi:https://doi.org/10.1016/j.gloplacha.2016.08.009

Mishra, S., Sarkar, U., Taraphder, S., Datta, S., Swain, D., Saikhom, R., . . . Laishram, M. (2017). Principal Component Analysis. *International Journal of Livestock Research*, 1. doi:10.5455/ijlr.20170415115235

Moreira de Sousa, L., Poggio, L., & Kempen, B. (2019). Comparison of FOSS4G Supported Equal-Area Projections Using Discrete Distortion Indicatrices. *ISPRS International Journal of Geo-Information, 8*(8). doi:10.3390/ijgi8080351

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *WIREs Data Mining and Knowledge Discovery, 2*(1), 86-97. doi:https://doi.org/10.1002/widm.53

O

Ozsahin, I., Uzun Ozsahin, D., Uzun, B., & Mustapha, M. T. (2021). Chapter 1 - Introduction. In I. Ozsahin, D. U. Ozsahin, & B. Uzun (Eds.), *Applications of Multi-Criteria Decision-Making Theories in Healthcare and Biomedical Engineering* (pp. 1-2): Academic Press.

P

Pagès, J. (2004). Analyse factorielle de données mixtes. *Revue de Statistique Appliquée, 52*(4), 93-111. Retrieved from http://www.numdam.org/item/RSA_2004__52_4_93_0/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Peduzzi, P., Dao, Q.-H., Herold, C., Diaz, A. M., Mouton, F., Nordbeck, O., . . . Widmer, B. (2002). *Global Risk And Vulnerability Index Trends per Year (GRAVITY) Phase II: Development, analysis and results*. Retrieved from https://archive-ouverte.unige.ch/unige:32353

R

Roy, B. (1968). Classement et choix en présence de points de vue multiples. In Revue française d'informatique et de recherche opérationnelle (Vol. 2, Issue 8, pp. 57–75). EDP Sciences. https://doi.org/10.1051/ro/196802v100571

S

Saaty, T. L. (1990). How to make a decision: The analytic hierarchy process. *European Journal of Operational Research, 48*(1), 9-26. doi:https://doi.org/10.1016/0377-2217(90)90057-I

Satta, A., Puddu, M., Venturini, S., & Giupponi, C. (2017). Assessment of coastal risks to climate change related impacts at the regional scale: The case of the Mediterranean region. *International Journal of Disaster Risk Reduction, 24*, 284-296. doi:https://doi.org/10.1016/j.ijdrr.2017.06.018

Sayre, R., Noble, S., Hamann, S., Smith, R., Wright, D., Breyer, S., . . . Reed, A. (2019). A new 30 meter resolution global shoreline vector and associated global islands database for the development of standardized ecological coastal units. *Journal of Operational Oceanography, 12*(sup2), S47-S56. doi:10.1080/1755876X.2018.1529714

Schmidtlein, M. C., Deutsch, R. C., Piegorsch, W. W., & Cutter, S. L. (2008). A Sensitivity Analysis of the Social Vulnerability Index. *Risk Analysis, 28*(4), 1099-1114. doi:https://doi.org/10.1111/j.1539-6924.2008.01072.x

Seekao, C., & Pharino, C. (2016). *Environ Earth Sci, 75*, 1.

Smit, B., & Wandel, J. (2006). Adaptation, adaptive capacity and vulnerability. *Global Environmental Change, 16*(3), 282-292. doi:10.1016/j.gloenvcha.2006.03.008

Spalding, M., Ruffo, S., Lacambra, C., Meliane, I., Hale, L., Shepard, C., & Beck, M. (2014). The role of ecosystems in coastal protection: Adapting to climate change and coastal hazards. *Ocean & Coastal Management, 90,* 50–57. doi:10.1016/j.ocecoaman.2013.09.007

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*, 72-101. doi:10.2307/1412159

Student, J., Kramer, M. R., & Steinmann, P. (2020). Simulating emerging coastal tourism vulnerabilities: an agent-based modelling approach. *Annals of Tourism Research, 85*. doi:10.1016/j.annals.2020.103034

Stock, J., & Watson, M. (2011). *Introduction to Econometrics (3rd edition)*: Addison Wesley Longman.

T

Tanim, A. H., Goharian, E., & Moradkhani, H. (2022). Integrated socio-environmental vulnerability assessment of coastal hazards using data-driven and multi-criteria analysis approaches. *Sci Rep, 12*(1), 11625. doi:10.1038/s41598-022-15237-z

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

U

UNEP. (2014). *The Adaptation Gap Report 2014*: United Nations Environment Programme (UNEP).

UNEP. (2021). *The Adaptation Gap Report 2021*: United Nations Environment Programme (UNEP).

V

Vafeidis, A. T., Nicholls, R. J., McFadden, L., Tol, R. S. J., Hinkel, J., Spencer, T., . . . Klein, R. J. T. (2008). A New Global Coastal Database for Impact and Vulnerability Analysis to Sea-Level Rise. *Journal of Coastal Research, 24*(4), 917-924. Retrieved from http://www.jstor.org/stable/40065185

Vousdoukas, M. I., Mentaschi, L., Hinkel, J., Ward, P. J., Mongelli, I., Ciscar, J. C., & Feyen, L. (2020). Economic motivation for raising coastal flood defenses in Europe. *Nat Commun, 11*(1), 2119. doi:10.1038/s41467-020-15665-3

W

Wang, Y., Fang, Z., Hong, H., & Peng, L. (2020). Flood susceptibility mapping using convolutional neural network frameworks. *Journal of Hydrology, 582*, 124482. doi:https://doi.org/10.1016/j.jhydrol.2019.124482

Wolff, C., Vafeidis, A. T., Muis, S., Lincke, D., Satta, A., Lionello, P., iJmenez, J. A, Hinkel, J. (2018). A Mediterranean coastal database for assessing the impacts of sea-level rise and associated hazards. *Sci Data, 5*, 180044. doi:10.1038/sdata.2018.44

Wong, P. P., Losada, I. J., Gattuso, J.-P., Hinkel, J., Khattabi, A., McInnes, K., Yoshiki, S., Sallenger, A. (2014). Coastal systems and low-lying areas. In (pp. 361-409).

Wu, T. (2021). Quantifying coastal flood vulnerability for climate adaptation policy using principal component analysis. *Ecological Indicators, 129*. doi:10.1016/j.ecolind.2021.108006

X

Xu, Q., Ding, C., Liu, J., & Luo, B. (2015). PCA-guided search for K-means. *Pattern Recognition Letters, 54*, 50-55. doi:https://doi.org/10.1016/j.patrec.2014.11.017

Y

Yager, R. R. (1988). On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on Systems, Man, & Cybernetics, 18*, 183-190. doi:10.1109/21.87068

Z

Zhang, Z., Hu, B., & Qiu, H. (2021). Comprehensive assessment of ecological risk in southwest Guangxi-Beibu bay based on DPSIR model and OWA-GIS. *Ecological Indicators, 132*, 108334. doi:https://doi.org/10.1016/j.ecolind.2021.108334

Appendix: Data Sources

1. Elevation

Amatulli, G., Domisch, S., Tuanmu, M.-N., Parmentier, B., Ranipeta, A., Malczyk, J., & Jetz, W. (2018). A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Scientific Data, 5(1), 180040. doi:10.1038/sdata.2018.40

A global digital elevation model - GTOP030 (123-99). (1999).

Those datasets have been employed in the definition of the low elevation coastal zones, following the procedure described in Merkens et al, 2016.

The first one (SRTM v. 4.1) is the most recent one and has been used for all the globe except for Antarctica and South Pole islands which were note reported.

To employ those dataset as elevation variable, I reprojected them from WGS84 to IGH with Qgis Gdal function Warpreproject

2. Distance from rivers

Lehner, B., & Grill, G. (2013). Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, *27*(15), 2171-2186. doi:https://doi.org/10.1002/hyp.9740

This data source was selected since, excluding the use of other providers which would have required an excessive amount of computational resources to extract data, such as OpenStreetMap project, it was judged to be the most comprehensive rivers dataset at the time of the analysis, with a spatial extent including all the globe.
Another major advantage was the included classification of the rivers under different criterias, which significantly facilitated the selection of the most relevant ones.

Once the geodatabase was downloaded, an r routine was employed to select all the rivers within the first three classes in the classical ordering system (ORD_CLAS < 3) and a long term average discharge at least equal to 10 m3/s (ORD_FLOW < 6).
Those criteria have been visually determined and manually tuned inspecting selected areas of the globe
The data was then converted from polygon to raster format data with terra::rasterize in R and then reprojected from WGS84 to IGH with nearest neighbour ,with Qgis Gdal function Warpreproject and nearest neighbor interpolation.
To compute the euclidean distance from the nearest river to each pixel Qgis Gdal Proximity tool was employed.

3. Presence of coastal ecosystems

1. Giri, C., E. Ochieng, L. L. Tieszen, Z. Zhu, A. Singh, T. Loveland, J. G. Masek, et al.2021 “Global Distribution of Mangroves USGS.” United Nations Environment Programme World Conservation Monitoring Centre (UNEP-WCMC). doi:10.34892/1411-W728

2. Mcowen, C., Weatherdon, L. V., van Bochove, J.-W., Sullivan, E., Blyth, S., Zockler, C., Stanwell-Smith, D., Kingston, N., Martin, C., Spalding, M., & Fletcher, S. (2021). Global Distribution of Saltmarsh (6.1) [Data set]. United Nations Environment Programme World Conservation Monitoring Centre (UNEP-WCMC). https://doi.org/10.34892/07VK-WS51

3. UNEP-WCMC, & Short, F. T. (2021). Global Distribution of Seagrasses (Version 7.1) [Data set]. United Nations Environment Programme World Conservation Monitoring Centre (UNEP-WCMC). https://doi.org/10.34892/X6R3-D211

4. UNEP-WCMC, WorldFish, World Resources Institute, & The Nature Conservancy. (2021). Global Distribution of Coral Reefs (Version 4.1) [Data set]. United Nations Environment Programme World Conservation Monitoring Centre (UNEP-WCMC). https://doi.org/10.34892/T2WK-5T34

As coastal ecosystems, mangroves, saltamarshes, seagrasses and coral reefs have been selected. This choice could be justified with these statement in IPCC, 2019: “major ‘protection’ benefits derived from the above-mentioned coastal ecosystems include wave attenuation and shoreline stabilisation” even if also “other ecosystems provide coastal protection […] but there is less understanding of the level of protection conferred by these other organisms and habitats”.

All the dataset presented here are certified from Unep and downloadable from its website. At the time of this analysis, alternative data sources with the same level of completeness were not available.

Those dataset, originally in shapefile format, have been rasterized with terra rasterize in a binary raster, representing absence or presence of any coastal ecosystem. Then , the data was reprojected from WGS84 to IGH with Qgis Gdal Warpreproject and nearest neighbor interpolation.

The coastal ecosystems which overlapped cells in coastal zones and the cells with a maximum distance of a kilometer, calculated with Qgis Gdal Proximity function, from those ecosystems are considered as protected.

There is no direct reference in the literature on a precise distance until those ecosystems exert a protective function, even if distance is frequently cited as a variable which is related to this phenomenon[47]. This could be also due to the fact that different ecosystems provide different degrees of protection and different protective function.

In absence of any information related to this issue, 1 km appeared as a reasonable buffer distance, but this step could be improved in the next analysis.

[47] For instance:
Spalding, M., Ruffo, S., Lacambra, C., Meliane, I., Hale, L., Shepard, C., & Beck, M. (2014). The role of ecosystems in coastal protection: Adapting to climate change and coastal hazards. Ocean & Coastal Management, 90, 50–57. doi:10.1016/j.ocecoaman.2013.09.007

4. Coastal erosion

Moosdorf, N., Cohen, S., & von Hagke, C. (2018). A global erodibility index to represent sediment production potential of different rock types. Applied Geography, 101, 36-44. doi:https://doi.org/10.1016/j.apgeog.2018.10.01

This dataset has been selected for its declared purpose of depicting the "landscape evolution" related to erosion.

While several datasets exist on (superficial) soil erodibility, mostly employed in agricultural studies[48], this one is able spatially represents the erosion potential of the rocks in the surface of the earth, from hard rocks to unconsolidated sediments.

Data was upscaled from an original spatial resolution of 0.1 degrees to a new spatial resolution of 0.008 degree, corresponding to 10 km at the equator. The it was reprojected from WGS84 to IGH. The reprojection algorith employed was the nearest neighbor interpolation, since it resulted the best option in terms of MAE, calculated as reported in section 3.1.
Both those operations were performed with Qgis Gdal Warpreproject.

Since a non-negligible amount of values was missing in the cells included in my coastal zones mask, I imputed those missing values with Qgis Gdal Fillnodata function, which uses a smoothing algorithm. I limited the cells being replaced to a maximum distance of 10 pixels from the nearest non-null cell, approximately corresponding to a radius of 10 km.

This operation was justified from a visual inspection to ensure that similar values are generally distributed in clusters next to each other, confirming in this case the first Tobler geography law.

5. Anthropogenic subsidence

Herrera-García, G., Ezquerro, P., Tomás, R., Béjar-Pizarro, M., López-Vinielles, J., Rossi, M., Ye, S. (2021). Mapping the global threat of land subsidence. American Association for the Advancement of Science (AAAS). https://doi.org/10.1126/science.abb8549

The only dataset publicly available on subsidence at the time of the analysis I am aware of is the aforementioned one.

The authors map subsidence from groundwater depletion susceptibility, which has been identified as one of the main causes of subsidence (Gambolati et al, 2005[49])
Combining this information with groundwater depletion probability they obtained a final spatial dataset with potential subsidence from groundwater depletion. I

[48] Such as, for example: Borrelli, P., Robinson, D. A., Fleischer, L. R., Lugato, E., Ballabio, C., Alewell, C., … Panagos, P. (2017). An assessment of the global impact of 21st century land use change on soil erosion. Springer Science and Business Media LLC. https://doi.org/10.1038/s41467-017-02142-7

[49] Gambolati, G., Teatini, P., & Ferronato, M. (2005). Anthropogenic Land Subsidence. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470848944.hsa164b

employed this data in the current version referred to 2010 and not its projection to 2040.

The data is categorical ordered in four classes, from 1 low to 6 high potential subsidence.

Since the data could be regarded as categorical I reprojected the map from WGS84 to IGH Qgis Gdal Warpreproject and nearest neighbor interpolation.

As in the case of subsidence, also in this case it was necessary to impute missing values. Since an algorithm preserving the original values was need, I employed Qgis Grass function r.fill.stats, filling the values with the mode of the nearest 10 cells if at least one cell was non-missing. Approximately equal to 10 km.

This operation is justified from the fact that the reasoning in the subsidence case holds true also for this variable.

6. Gross domestic product

Kummu, M., Taka, M., & Guillaume, J. H. A. (2019). Data from: Gridded global datasets for Gross Domestic Product and Human Development Index over 1990-2015 (Version 2) [Data set]. Dryad. https://doi.org/10.5061/DRYAD.DK1J0

To my knowledge, at the time of the analysis other two datasets could have been employed to include gridded Gdp projections in my analysis:

Wang, T., & Sun, F. (2022). Global gridded GDP data set consistent with the shared socioeconomic pathways. Scientific Data, 9(1), 221. doi:10.1038/s41597-022-01300-x

Murakami, D., & Yamagata, Y. (2019). Estimation of Gridded Population and GDP Scenarios with Spatially Explicit Statistical Downscaling. MDPI AG. https://doi.org/10.3390/su11072106

This source directly disaggregate official statistics on Gdp per capita obtained at the finer possible level, using the population count as the unique downscaling variable, while the other sources employ more sophisticated methods such as projection according to night-time light (Wang and Sun, 2022) or more auxiliary variables related to Gdp (Murakami and Yamagata, 2022).

The latest year available is 2015, thus there are not future projections according to the SSPs as in the other sources, but this is fine since this analysis is a current state analysis.

In Wang and Sun, 2022, the resolution is higher[50] compared to Kummu et al, 2019 but the representation of the coasts is rougher than in the source I selected, resulting in a higher amount of cells presenting missing values. The same is true for Murakami and Yamagata, 2022, even if the original resolution in this case is the same as in Kummu et al., 2019.

[50] 1 km vs 10 km

Other advantage of this dataset include the fact that the values are presented in purchasing power parity and therefore directly comparable across different regions of the world and the fact that the same publications present gridded data on HDI too. Since I also am using including this variable, using the same source for gdp could improve the coherence of the analysis.

Even if all these sources have their advantage and disadvantages, everything considered Kummu et all appears as the data source most suited for my purposes.

Before reprojecting I changed the resolution from 10 to 1 km attributing to each cell on the grid one tenth of the original value at 10 km resolution, since the Gdp in the cell is reported as a total and not per capita value.

Once selected this option, I used Qgis Gdal Warpreproject with nearest neighbor algorithm to reproject data from WGS84 to IGH. This option was chosen since it minimized MAE.

7. General state of development

Kummu, M., Taka, M., & Guillaume, J. H. A. (2019). Data from: Gridded global datasets for Gross Domestic Product and Human Development Index over 1990-2015 (Version 2) [Data set]. Dryad. https://doi.org/10.5061/DRYAD.DK1J0

As previously mentioned, the source is the same as for Gdp. In this case there is no variable available to disaggregate the values in lower, 10 km cells resolution, hence the data is available at the lowest possible administrative subdivision for each area. However, the data is available in raster format, with a resolution of 10 km[51].

It was not possible to find publicly available data on HDI at a finer spatial level.

I changed the resolution from 10 to 1 km and reprojected the map with Qgis Gdal Warpreproject and nearest neighbor interpolation algorithm

8. Presence of vulnerable social groups

Center For International Earth Science Information Network-CIESIN-Columbia University. (2018). Gridded Population of the World, Version 4 (GPWv4): Basic Demographic Characteristics, Revision 11 [Data set]. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H46M34XX

This dataset represents, to my best knowledge at the time of the analysis, the only source publicly available which presents gridded global level data on basic demographic groups, like age cohorts and gender subdivisions.
This data is derived from official sources / census and is available for the year 2010 with 1 km resolution.

[51] Cells within the same administrative area share exactly the same value, data is usually available at an administrative level lower than the national one.

I selected the count of people with an age of 14 or younger to represent the young cohort, the count of people older than 65 years to represent the elderly and the count of female population

I then computed the proportions of those groups on the total population reported the same cell in 2010, assuming this proportion until 2020, year for which I downloaded data on total population count from the same source, as I am going to report in the next part of this section.

The data was then reprojected from WGS84 to IGH with Qgis Gdal Warpreproject and nearest neighbor option, which resulted the most effective in terms of MAE in this case.

9. Economic inequality

The World Bank, World Development Indicators (2021). Gini index [Data set]. Retrieved from https://databank.worldbank.org/reports.aspx?source=2&series=SI.POV.GINI

10. Gender inequality

Unep (2021). GENDER INEQUALITY INDEX (GII) [Data set]. Retrieved from https://hdr.undp.org/sites/default/files/data/2020/2020_Statistical_Annex_Table_5.xlsx

11. Effectiveness of governance

The World Bank, Worldwide governance indicators (2021). Government effectiveness [Data set]. Retrieved from https://databank.worldbank.org/source/worldwide-governance-indicators

The government effectiveness indicator compute by the world bank appeared as the most suited to represent the quality of the government in a territory, combining several dimensions associated with the phenomena.
This index is an aggregation of several survey based sources and hence relies on the perceptions of the citizens and organizations.

12. Effectiveness of the Juridical system.

The World Bank, Worldwide governance indicators (2020). Rule of law [Data set]. Retrieved from https://databank.worldbank.org/source/worldwide-governance-indicators

Coming from the same set of indicators as government effectiveness, this index is built with the same principles as the previous variable.
In this case the purpose is “to measure the perceptions of the extent to which agents have confidence in and abide by the rules of society” (World Bank[52]).
Given how this indicator is constructed, I can reasonably assume that it could well represent the role of justice in adaptation to climate change, as described in the already cited IPCC reports.

[52] World Bank official webisite, url: https://tcdata360.worldbank.org/indicators/hf5cdd4dc?country=BRA&indicator=370&viz=line_chart&years=1996,2020

13. Tourism

The World Bank, World Development Indicators (2021). International tourism, number of arrivals [Data set]. Retrieved from https://databank.worldbank.org/reports.aspx?source=2&series=ST.INT.ARVL&country=

The World Bank, World Development Indicators (2021). Population, total [Data set]. Retrieved from https://databank.worldbank.org/reports.aspx?source=2&series=SP.POP.TOTL&country=

All the variables from 8 to 11 were not available in spatial format at the time of the analysis or the few existing spatial sources were discarded for an overall lack of quality[53]. Notheless, this informations were available at best on national level. It's however noteworthy that assuming the majority of this metrics as constant at a national level could be reasonable.

Given this premises, I had to adopt a procedure to spatialize those data.
With Python and Geopandas library I have been able to link the first level administrative divisions (national) geometries to the related value of the non-spatial indicators, through the official country code.
After that, I had to ensure that the values retained were referred to the latest year available for each country.
In this way I obtained a complete spatial vector dataset, in shapefile format.

For the territories for which I was able to perform this operation, but the value of the indicator was not available in the original non-spatial dataset, I imputed the missing value with the average value of the indicator in the same geographic area, according to the world bank classification. Taking this step was only possible for territories with an associated official code, which is, again, related to the world bank classification.

Areas in the GADM database without an associated code were then inspected to manually correct possible errors: to French Guyana the same values, when the data was missing, I assigned the same value of France and to Western Sahara and Northern Cyprus, disputed areas, I manually imputed the values of the related geographic area.

To the other areas left with missing values, mostly islands and territories with a very limited spatial extension, negligible as compared to the global areas considered, I imputed the total average values of the territories that were originally missing before the data imputation on the first step.

Concerning tourism, the data processing steps differed in the fact that I had to compute the ratio of international tourism arrivals with the total population of the country, as derived at national level from the world bank, and from the missing values imputation strategy: in this case I simply replaced missing value with zeros,

[53] In the case of tourism a spatial dataset publicly available was discarded for the lack of spatial resolution and insufficient spatial coverage, with several missing values: Adamiak, C., & Szyda, B. (2021). Combining Conventional Statistics and Big Data to Map Global Tourism Destinations Before COVID-19. SAGE Publications. https://doi.org/10.1177/00472875211051418. For Gini index a spatial-based and more detailed dataset has been published in the SEDAC data portal, but is still under review

assuming that for territories with missing statistics the amount of tourists is negligible.

After the missing values imputation, I employed terra rasterize function to convert each of the dataset in raster format and then I projected the data from WGS84 to IGH with Qgis Gdal Warpreproject function and nearest neighbor algorithm.

14. Population count

Center For International Earth Science Information Network-CIESIN-Columbia University. (2018). Gridded Population of the World, Version 4 (GPWv4): Population Count, Revision 11 [Data set]. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4JW8BX5

This dataset consists in a proportional projection of population count data obtained at the finest possible administrative level, exploiting data on approximately 13.5 million of administrative units it represented the gridded allocation of population with the most accurate starting data source.

Differently from other sources[54], this data is only suited for a current level analysis since the latest year available, employed here is 2020, but it has the advantage of actually using official data at a particularly fine level on population and not recurring to sophisticated downscaling algorithms such as the ones based on night-time light.

Another advantage of this source is the fact that it is the only one presenting data on demographic groups, as described in 7. Vulnerable social groups.

Once chosen this source, I had to reproject the map from WGS84 to IGH with Qgis Gdal Warpreproject and nearest neighbor algorithm, which resulted in the lowest interpolation MAE.

15. Percentage of area covered by urban settlements

Pesaresi, M., & Politis, P. (2022). GHS-BUILT-S R2022A - GHS built-up surface grid, derived from Sentinel2 composite and Landsat, multitemporal (1975-2030) [Data set]. European Commission, Joint Research Centre (JRC). http://doi.org/10.2905/D07D81B4-7680-4D28-B896-583745C27085

This dataset present information on the estimated number of squared meters allocate to residential and non-residential built-up surfaces. This data is derived from satellite images at an extremely fine global resolution of 10 meters and then aggregated into coarser resolutions. I downloaded the 1 km resolution version of the dataset.

Although information on urban land cover is reported in other datasets, such as the one on land cover from ESA CCI that I am discussing later in 16. Share of agricultural land cover, this is the only one that could be used to derive information on the share of urban land cover at the most precise level[55].

[54] Such as, for instance, the already cited Merkens et al, 2016

[55] In other land cover datasets the cell is usually classified in urban or not.

The map was then reprojected from Mollweide projection to IGH with Qgis Gdal Warpreproject and nearest neighbor algorithm, which resulted in the lowest interpolation MAE.

16. Density of main roads and presence of transportation nodes

Meijer, J., Huijbregts, M., Schotten, K., & Schipper, A. (2018). Global patterns of current and future road infrastructure. Environmental Research Letters, 13. doi:10.1088/1748-9326/aabd42

This dataset, cited among the others in the Food and agriculture organisation of the United Nations (FAO) and World Bank data catalogs, was chosen for this analysis since it was judged as the most complete dataset already available on the topic.

An alternative could have been to extract data from other providers such as OpenStreetMap, but I discarded this operation since it resulted prohibitively computationally expensive.

The roads, represented in line geometries, are classified according to their relevance: I extracted the two principal classes only, highways and other main roads.

Then, I produced a routine in R to count the number of line segments crossing each cell in a raster file sharing the same properties as the baseline coastal zones mask raster. I assumed that the cells crossed by a great number of streets segments are likely to be transportation nodes.

The map was then reprojected from WGS84 to IGH with Qgis Gdal Warpreproject and nearest neighbor algorithm.

In this case, given the peculiar spatial distribution of the values, concentrated only on the lines representing the roads, and the fact that data is of integer type, the reprojection with cubic splines was not considered to avoid, since the smoothing effect is not well suited to represent this phenomenon.

17. Agricultural land cover

Copernicus Climate Change Service. (2019). Land cover classification gridded maps from 1992 to present derived from satellite observations [Data set]. ECMWF. https://doi.org/10.24381/CDS.006F2C9A

This dataset was selected since it was the only one I was aware at the time of the analysis presenting data on agricultural land covers at an adequate spatial resolution for the most recent years.

I selected the data for 2020 as it was the most recent year available, and then I classified the cells in a binary variable according to their agricultural or non-agricultural destination.

After this step I had to aggregate the values from approximately 0.002 degrees to a 0.008 degree resolution (roughly 1 km at the equator), with Qgis function Save layer as… and changing the resolution with mean as aggregation method.

In this way, the resulting cells now present an approximation of the fraction of agricultural land cover.

Then I had to reproject the map from WGS84 to IGH with Qgis Gdal Warpreproject and nearest neighbor algorithm, which resulted in the lowest interpolation MAE.

18. Presence of natural heritage sites

UNEP-WCMC & IUCN (2022). Protected Planet: The World Database on Protected Areas (WDPA) [data set]. Cambridge, UK: UNEP-WCMC and IUCN. Retrieved from www.protectedplanet.net

UNEP-WCMC & IUCN (2022). The World Database on Other Effective Area-based Conservation Measures (WD-OECM [data set]. Cambridge, UK: UNEP-WCMC and IUCN. Retrieved from www.protectedplanet.net

Within data sources it is possible to find all the natural protected areas of the world represented in vector format. The first dataset contains the areas for which the preservation of biodiversity is a primary objective, while the second one contains the areas for which this function is a result of other objectives or a secondary objective. Those areas include, for instance, UNESCO natural and mixed world heritage sites.

Therefore, I can reasonably assume that the institution of this kind of area is strongly related to presence the natural heritage variable cited in the mentioned IPCC reports.

This data is periodically updated, I employed the september 2022 version

Since buffering operations were required in the workflow, employing the data in a equal are projection was required in this case. Hence, as a first step I had to reproject the vector data from WGS84 to IGH with Qgis native function Reproject vector

After this operation, I had to convert the (minority) of areas in this datasets associated with a point format to a polygon. I followed the procedure described in the official documentation of the datasets: if an area in meters was reported on the point record, I represented these territories with a circular buffer proportional to the reported areas, otherwise I discarded the record.

Then, I discarded the marine protected areas, since I am focusing on the coastal mainland and this significantly improved the computational requirements of the following.

After that, I rasterized the areas with R terra rasterize

Table 12: Mean absolute error, reprojection algorithms

Variables	**Nearest neighbor**	**Cubic spline**
elevation	12.8575	15.6048
erodibility	0.0105	0.0990
GDP_PPP_2015	26762.6000	77864.3000
female_tot_pop_ratio	0.0035	0.0059
old_tot_pop_ratio	0.0025	0.0039
young_tot_pop_ratio	0.0022	0.0036
total_pop_2020	3.6584	6.8613
urban_areas	924.6440	1.325.7400
land_cov_agr	0.0275	0.0376

Ringrazio i miei genitori e tutta la mia famiglia per il sostegno e l'immensa pazienza.
Ringrazio in generale le persone che sono sempre state al mio fianco.
Un ringraziamento particolare a Michele Zen ed al prof. Giupponi per i preziosi consigli.