# Introduction

Intraspecific genetic diversity can be beneficial in allowing species to adapt to new environmental conditions and contributes to population viability, community and ecosystem functioning, and nature’s contributions to people (Razgour et al 2019; Palumbi et al 2019; Des Roches 2019, 2021). As there is mounting evidence of significant declines in intraspecific genetic diversity in wild species (Exposito-Alonso et al. 2022; Leigh et al. 2019), the importance of preserving intraspecific genetic diversity is now recognized in applied conservation. Proposed targets of conservation for genetic diversity range between 90% and 99% of a species extant genetic diversity (Hoban et al 2023). According to Frankham (2022), retaining at least 99% of extant, within-population genetic diversity would be needed to prevent a sharp rise in inbreeding and a reduction in individual fitness, avoiding threats to population and species viability.

Conservation areas (CAs) are a very valuable tool to protect biodiversity when they are properly planned, designed and managed (Maxwell et al 2020). It is therefore important to have clear frameworks for evaluating to what extent a system of CAs protect the genetic diversity of species and to devise methods to integrate the targets for genetic diversity conservation into the spatial planning of protected areas. To this end, it is possible to leverage on recent advances into integration of genetic data into systematic conservation planning (Andrello et al 2022; Andrello et al 2023; Nielsen et al 2022). Systematic conservation planning, and in particular spatial conservation prioritization (SCP), is a decision framework to allocate scarce resources to spatial conservation actions, such as the creation of protected areas, following principles of comprehensiveness, adequacy and efficiency (Kukkala & Moilanen 2013; Margules and Sarkar 2007; Margules and Pressey 2000; Moilanen *et al.* 2009b). Comprehensiveness means that the PA network should sample the full range of biodiversity (ref); adequacy means that biodiversity should persist in the long term; and efficiency means that the set of identified protected areas must respond to budget constraints (Kukkala & Moilanen 2013). In the most common SCP applications, species or habitat types are the conservation features (i.e. entities that can be mapped in the landscape). For example data on the presence/absence of species in planning units (PUs) in the landscape are used to identify the set of PUs fulfilling explicit conservation objectives, such as to maximize the representation of species in the set of PUs selected for protection. The principle of adequacy is usually pursued by setting amount representation targets for species presences, usually in the form of proportion of their distribution ranges, because protecting multiple occurrences of a species reduces the risk of species extinction following local losses. A species is considered adequately represented if the proportion of its distribution range included in the selected PUs is above its amount representation target.

Genetic diversity has only recently been considered in evaluations of CAs (Munguia-Vega et al 2015 GEC; Vasconcelos et al 2018; Benestan et al 2023 Ecography) and in the planning of new CA systems through SCP (Andrello et al 2022; Nielsen et al 2022; Andrello et al under review). Various indicators of genetic diversity have been discussed within scientific reflections on essential biodiversity variables. In particular, four essential variables have been proposed to describe the genetic composition of species: allelic richness, heterozygosity, number of genetically differentiated populations and effective population size (Hoban et al 2022). These variables may refer to different levels of organization of genetic composition. The individual level includes variables that measure genetic diversity within individuals, such as observed heterozygosity or runs of homozygosity. The population level includes variables related to genetic diversity within populations, such as allelic richness, expected heterozygosity or effective population size. Higher levels are groups of populations within a region or entire species ranges, for which variables could be allelic richness at the species level, the number of distinct genetic lineages, or the effective population size of the species. SCP studies have used some of these variables, such as allelic richness at the regional level (Paz-Vinas et al 2018), the number of genetic lineages in a region (Vasconcelos et al 2012, 2018) or the expected heterozygosity at the PU level (Nielsen et al 2017; Xuereb et al 2021). These variables were used to define minimum-set SCP problems to maximize the representation of genetic diversity within a system of CAs.

However, these variables are difficult to use for species that exhibit spatial genetic variation without genetically differentiated populations, such as species with organisms distributed continuously in space, species showing patterns of clinal variation in allele frequencies or patterns of genetic patchiness (Selkoe et al 2010). How to choose genetic diversity variables in these cases? Variables defined at the population level are difficult to use because distinct populations cannot be identified. Variables defined at a higher level (groups of populations or entire species ranges) would lose information about the spatial structure of genetic diversity. In these cases, genetic diversity must therefore be described by metrics measured at the individual level, but it is necessary at the same time to find a method for using metrics measured at the individual level in spatial planning approaches whose planning units are spatial portions of land.

In this paper, we explore how to use PU-level genetic metrics obtained through PCA on individual-level genetic data. In particular, we focus on the parameterization effects of two approaches to evaluate a protected area system and to indentify an optimal protected area system with respect to the representation of genetic diversity. The first approach has been called by us the split-taxon approach and was initially proposed by Beger et al (2014). The second approach is implemented in the raptr program (Hanson et al 2017). Approaches to including genetic diversity in spatial conservation prioritization

## Use of PCA to define PUs-level variables from individual genetic data

One approach to preserve the information about genetic structure at multiple loci is principal component analysis (PCA). PCA and other types of statistical sorting (PCoA, sPCA ) allow the information contained in the genetic data to be sorted by defining new variables (the principal components) that describe the positioning of each individual or groups of individuals relative to the total genetic variation contained in the data set. In SCP, after calculating the scores for each individual or sample, one need to apply a spatialization step to predict PCA scores in unsampled PUs. As samples used in genetic analyses are usually taken from a small portion of the range of occurrence of a species, the objective of the spatialization step is to predict the value of metrics over the entire range of occurrence. This usually involves fitting a statistical model that links metrics to spatial coordinates and using it to predict the values of metrics across space. Several statistical models are available (nearest neighbor, kriging, idw; Wagner & Fortin ).

Once genetic diversity values have been obtained across the range of a species, the selection of PUs to protect can be done using prioritzr or raptr.

## Selection with *prioritzr*

Prioritizr is an R package that can be used to solve various SCP problems (minimum set, maximal coverage etc) involving conservation features, which are discrete entities that have an occurrence (0/1) or a density in each PUs. To reconduct continuous genetic metrics such as PCA score to discrete conservation features one can use a split-taxon (ST) approach.

The split-taxon (ST) approach consists in dividing the set of PUs in which a taxon occurs into classes based on the genetic metric of interest. The taxon can be a species or a grouping below the species level (lineage, evolutionarily significant unit or management unit; Funk et al 2012) that exhibits intra-taxon variation measured by the continuous genetic metric of interest, e.g. the scores of a PCA of allelic counts. For example, one can split a species into three classes of low scores, medium and high PCA scores. Classes can be defined by several criteria. First, one can define classes at equal intervals along the range of variation of the variable: this makes classification in terms of high or low values intuitive but could result in a different number of PUs in each class. Second, one can define classes using quantiles of the distribution, which has the advantage of creating classes with an equal number of PUs but could represent ranges of variation of different magnitude. Third, more sophisticated statistical clustering techniques, such as the k-means algorithm, can be used to define the optimal number of classes and the partition of PUs into them.

To optimize a selection of candidate PUs, one can define spatial targets of representation for each genetic class. Usually, a taxon is considered adequately protected when protected areas cover a portion of its range greater than a target. This amount target amount can be a decreasing function of a taxon's range. For example, target can decrease from 100% for species with less than 1,000 km2 of range to 10% for those with more than 250,000 km2 of range (see citation in Hanson et al 2018). Similarly, representation targets can be defined for each class. The same targets can be used for all classes (e.g., Hanson et al 2018), or different targets can be defined for different classes. This type of target is called an amount or representation target, and it indicates the minimum number of PUs needed to ensure the representation of a given class of genetic diversity. A set of PUs is considered representative when the targets for all classes are reached. In the presence of many classes and/or taxa, representativeness can also be calculated as the number of classes that have reached their target amount over the set of taxa.

This split-taxon approach to representation of genetic diversity makes it easy to define spatial prioritization problems with mathematical form that are identical to those used to ensure species representation. In the minimum set problem, the objective is to minimize the total conservation cost needed to meet the representation targets for all classes. In the maximal coverage problem, the objective is to maximize the classes that achieve their representation targets under the condition that the total conservation cost does not exceed a predefined budget. These mathematical problems can be solved by heuristic, semi-heuristic techniques or algorithms that guarantee optimality of solutions, such as those implemented in Marxan and prioritizr.

## Selection with *raptr*

As presented by Hanson et al (2017), raptr combines the optimal facility location approach developed in the field of operations research with the environmental diversity approach proposed by Faith (1992) to maximize the representation of phylogenetic or environmental diversity. The optimal facility location problem is to find an optimal set of PUs for the construction of new facilities, e.g., schools or hospitals, while minimizing the total distance between them and the demand points, e.g., population centers. In this case, demand points are physical entities located in the two-dimensional geographic space defined by the two variables longitude and latitude. The spatial planning problem can be formulated as: “to find PUs for new facilities so that the geographic distance between them and demand points is minimal”. Such spatial planning problems can be solved using software like the Gurobi solver (ref).

[However, more often, ] spatial planning of protected areas aims to locate PAs to cover a range of environmental conditions representative of the habitats to be protected. In this case, the facility location approach is used on environmental rather than geographical spaces, e.g. using a two-dimensional space defined by temperature and precipitation, and mapping PUs in this environmental space. Demand points are pairs of temperature and precipitation values of conservation interest, such as rare or extreme environments, or environments that are suitable for the conservation of locally-adapted populations. In this case, the spatial planning problem is formulated as: “to find PUs to protect so that the environmental distance between them and the demand points is minimal”. For example, in an application of raptr on four bird species in Northwest Australia, Hanson et al (2018) defined an environmental space using the first two axes of a PCA of 19 bioclimatic variables. They then placed a demand point in each PU where the species were present because they wanted to represent the ecological niche occupied by each species. The Gurobi solver was then used to minimize the environmental distance between the PUs selected for protection and the demand points.

To represent genetic diversity, demand points can be placed in the multidimensional space defined by the axes of the PCA. The spatial planning problem is formulated as: “to find PUs for a set of protected areas such that the genetic distance between them and the demand points is minimal”. For example, in Hanson et al (2018), raptr was used on a genetic space defined by the first two axes of a NMDS conducted on AFLP data on a species of plant in the European alps. By minimizing the distance between selected PUs and demand points, the genetic variation of the species could be represented in a minimum set of PUs. **Figure\_example\_raptr** show how to create a two-dimensional space defined by the first two axes of a PCA for an imaginary set of 50 planning units.

Immagine che contiene diagramma, schermata, testo, linea

Descrizione generata automaticamente

**Figure\_example\_raptr.** Definition of genetic spaces, demand points and the portion of conserved genetic diversity S2 (see equation (1)) on a hypothetical landscape with 50 square planning units (PUs). Panels (a) and (b) show the distribution of PUs in physical space and the value of two continuous genetic metrics, which could be the score on the axes of a PCA of allele counts. Two possible selections of three PUs to be retained (green border) are shown in (c) and (e). For each, panels (d) and (f) show the two-dimensional space defined by the two genetic metrics with PUs (gray dots; with green border if selected), three demand points (black dots) and the distance between each demand point and a PU (blue segments, corresponding to the distances dij in equation (1)).

To evaluate a selection of PUs in terms of genetic representativeness, the portion of conserved diversity is calculated with a ratio of sums of squared distances:

Where *dij* is the distance between demand point *i* and PU *j*, *δi* is the distance between demand point *i* and the centroid of demand points, and sums are taken over all demand points and all PUs in the selection. The *S*2 metric is analogous to the sum-of-squares subdivision in the *k*-means algorithm: the demand points in raptr correspond to the observations in the *k*-means algorithm; the PUs in raptr correspond to the group averages in the k-means; the denominator of the fraction in Eq. (1) corresponds to the total sum of squares; and the numerator corresponds to the within sum of squares. Intuitively, a set of demand points with more diverse genetic values will have a larger total sum of squares than a more compact set of demand points. A set of PUs closer to the demand points (as that shown in panel (c) and (d) of **Figure\_example\_raptr**) will have a lower within sum of squares, and better represent the diversity of demand points, than a set of PUs further from the demand points (as that shown in panel (e) and (f) of **Figure\_example\_raptr**). Low values of WSS correspond to high values of *S*2, which can be interpreted as the portion of genetic diversity conserved by the selected PUs. *S*2 can be used to define a protection target τ for genetic diversity by imposing the mathematical constraint *S*2 > τ. Finally, the distribution of demand points in relation to the distribution of PUs determine the maximum achievable genetic target max(*S2*), which is defined as the value of *S2* when all PUs are selected.

## Methodological choices likely to affect the selection of protected areas

Several methodological choices are likely to have effects on the results and performance of SCP, including the computation time, the maximum achievable genetic targets, the final selection of PUs, and their total conservation cost. The first choice, common to both prioritization methods, is how many axes of PCA to consider in the prioritization. In published studies, which used only raptr, only two axes have been considered (Hanson et al 2017, 2018; verify), but one criterion might be to choose a number of axes that represents a certain fraction of genetic diversity (looking at the eigenvalues associated with each axis). The second choice is whether to consider each axis separately or to combine them to define conservation features (in prioritizr) or genetic spaces (in raptr). The published examples (Hanson et al 2017, 2018; verify) chose to combine the two axes to define a two-dimensional genetic space, but there has not been a study showing how results change with the number of PCA axes included in the prioritization.

In addition, each of prioritizr and raptr involve specific methodological choices. With prioritizr, when considering individual PCA axes as distinct conservation features, one must choose how many classes to subdivide PUs scores into and how to subdivide them. The more classes, the less distortion caused by mapping continuous values into discrete classes, but the number of conservation features and thus the computational demands of the problem will increase. There are also various methods to subdivide continuous values into discrete classes. Similarly, when using multiple PCA axes to define conservation features, one must choose the number of clusters and the clustering algorithm.

With raptr, the main choices are how many demand points to use to describe the genetic spaces and how to distribute the demand points. These choices are analogous to choosing the number of discrete classes and how to partition the values into discrete classes with prioritizr. In fact, using a larger number of demand points results in a more faithful mapping of genetic spaces with discrete points, but involves computationally more complex problems. The distribution of demand points can also be done as in the prioritzr case, that is, by dividing the continuous values into classes (in the one-dimensional case) or into clusters (in the multidimensional case) and generating a demand point at the median point of each class. In the multidimensional case, raptr offers the option to generate demand points using various hypervolume calculation methods (ref).

## Purpose of this work

Here, we aim at integrating genetic conservation targets in spatial conservation prioritization for two fish species in the Mediterranean sea, the white seabream *Diplodus sargus* and the red mullet *Mullus surmuletus*. These are the only two species that have been sampled throughout most of their range of presence in the Mediterranean sea following a regular sampling grid, a fundamental prerequisite to map their genetic diversity in space (Dalongeville et al 2018 Evol Appl; Boulanger et al 2020 Ecography). Moreover, both species were genotyped via a Genotyping-by-Sequencing approach providing a high number of Single Nucleotide Polymorphisms (SNPs). This allowed gaining insights into the putatively neutral and adaptive components of genetic diversity, which is important as spatial conservation priority can be significantly different for neutral and adaptive loci (Xuereb et al 2021; Hanson et al 2021). Through different prioritization scenarios, we aim at answering the following questions:

How do the methodological choices affect the performance and the results of the planning process?

Where are the priority areas for conserving genetic diversity for these two species? (comparison across scenarios)

What proportion of genetic diversity is protected by the current system of MRs in the Mediterranean sea? How could we extend the current system of MRs to reach the targets of protection of genetic diversity?

What proportion of genetic diversity can be conserved with SCP based on species occurrence only without genetic data?

# Methods

## Study area and species

The study area encompassed the coastal areas (between 0 and 200 m depth) of the Mediterranean basin. We defined squared planning units (PUs) of 10-km size in Albers Equal Area projected coordinate reference systems. This projection ensures that all PUs have the same surface area (100 km2). Bathymetry was retrieved from ETOPO1 (ref). The final region had 5216 PUs. For comparing different methods to include genetic data, we used a unitary cost for all PUs so that the selection of PUs was driven only by the methods used to include the genetic information. For the final planning exercise, we integrated realistic conservation costs reflecting opportunity costs for fishing and aquaculture (Mazor et al 2014).

We did not attempt to estimate a conservation cost for each PU and instead set PU cost proportional to their surface area, as often done in these cases. Because PUs have all the same surface area in the chosen coordinate reference system, this led to setting a unitary cost for all PUs. Therefore, our results cannot be taken as realistic conservation prioritization; however, setting unitary cost facilitates the interpretation of the SCP solutions as they are driven only by the distribution of conservation features, amount representation targets and genetic conservation targets.

The white seabream *D. sargus* and striped red mullet *M. surmuletus* are coastal fishes distributed over all the Mediterranean basin. *D. sargus* inhabits rocky reefs and seagrass beds, whereas *M. surmuletus* mostly occupies sandy and rocky habitats (Froese & Pauly, 2019). [Details on importance for fishing and ecology (urchin-seabream relationships)]. The geographic distribution of the two species was taken from the Fishmed database (Albouy et al 2015), which maps the range of Mediterranean fishes at 1/10th degree and resulted in *D. sargus* and *M. surmuletus* present in 2266 and 3626 PUs, respectively.

## Genetic data

We used the intraspecific genetic data for *D. sargus* and *M. surmuletus* presented in Boulanger et al (2022). Briefly, 297 *D. sargus* and 526 *M. surmuletus* individuals were collected in 2014 from 59 and 64 sampling cells, respectively, spanning most of coastal areas of the Mediterranean basin (show figure). Individuals were genotyped via a Genotyping by Sequencing approach using restriction enzymes Pst1/Bfa1 (for *D. sargus*) and ApeKI (Elshire et al., 2011; for *M. surmuletus*) using the Illumina HiSeq 2500 (100 bp, single-end reads) at Cornell University. Raw sequences were filtered, trimmed, demultiplexed and mapped on the published reference genomes (Fietz et al., 2020). Variant calling and further filtering steps resulted in 8206 and 2794 SNPs for *D. sargus* and *M. surmuletus*, respectively. Loci that were more differentiated than under a neutral model (outlier) were identified using PCAdapt (Luu et al., 2017). For the non-outlier dataset, only SNPs in Hardy–Weinberg Equilibrium (HWE) were kept (Beaumont & Nichols, 1996). These resulted in final datasets of 7655 neutral SNPs and 413 outlier SNPs for *D. sargus*, and 2462 neutral SNPs and 291 outlier SNPs for *M. surmuletus*. We combined the neutral and outlier loci for each species, obtaining final datasets of 8068 and 2753 loci, respectively. Further details on the generation of these datasets can be found in Boulanger et al (2022).

We then applied a principal component analysis to the matrix of allele counts per sampling site using the ‘dudi.pca’ function in the R package *ade4* v. 1.7-22. We retained 17 (for *D. sargus*) and 26 (for *M. surmuletus*) PCA axes as they explained 80% of the variation. We then interpolated the score of the sampling sites over the 5216 PUs by fitting a nearest-neighbor model with the function ‘gstat’ in the R package *gstat* v. 2.1-1 (Gräler et al 2016) and using it to predict PCA scores over a raster with the function ‘interpolate’ in the R package *terra* v. 1.7-29.

## Definition and solution of planning problems

We defined planning problems in prioritizr and raptr separately for the two species, using different options to include the genetic information stored in the rasters of interpolated PCA scores.

### Prioritizr

In prioritizr, we first explored planning problems defined using single PCA axes as conservation features. We explored four different methods to subdivide the single-axis PCA scores into discrete classes: equal interval, quantile, standard deviation, and natural breaks (all as implemented in the R package *rgeoda* v. 0.0.10-2; Li & Anselin, 2023). Equal interval divides the range of values into equal intervals; quantile defines intervals so that each of them contains the same number of records; standard deviation subdivides the range of values into six intervals defined by the mean and standard deviation of the records [i.e. min(*x*), m-2s, m-s, m, m+s, m+2s, max(*x*)]; natural breaks groups observations such that the within-group homogeneity is maximized. We explored each of these classification methods using three numbers of classes (3, 6, and 12) for each PCA axis, except for standard deviation that was run with 6 classes only.

We then explored planning problems where we combined multiple PCA axes to define conservation features. This was done by running a *k*-means clustering algorithm (function ‘kmeans’ from the R package *stats* v. 4.3.0) on the PCA scores of all axes to group sampling sites into a discrete number of clusters. We ran several *k*-means algorithms using *k* = 1 to 20 clusters and selected the number of clustersthat minimized the within sum-of-squares, i.e. *k* = 5 for *D. sargus* and *k* = 7 for *M. surmuletus*. The identification of the optimal number of clusters in *k*-means can be subjective, because all criteria (e.g. within sum-of-squares or BIC) tend to decrease artificially as the number of clusters increase. Therefore, we also considered *k* = 2 for *D. sargus* and *k* = 3 for *M. surmuletus*, following results of Boulanger et al (2022).

We thus defined a total of 12 planning problems (10 using single PCA axes and 2 using multiple PCA axes) with prioritizr for both species (**Table**).

In prioritizr, we assigned a 15% representation target to all conservation features and defined planning problems under the minimum set formulation, i.e. finding the selection of PU that maximized the targets met while minimizing conservation costs. We used the Gurobi 9.5.2 optimization tool (REF) and generated a portfolio of 100 solutions (using ‘add\_gap\_portfolio’ in prioritizr) within 2% optimality.

### Raptr

In raptr, we first used each PCA axis as a distinct, one-dimensional genetic space. We used the same classification methods used for prioritizr (equal interval, quantile, standard deviation and natural breaks) to place demand points in the genetic spaces, except that, in the case of raptr, these methods were not used to define breaks in the interval of values but to find the midpoint of each interval. We placed one demand point in the midpoint of each interval and assigned it a weight proportional to the number of PUs included in that interval; if an interval had zero PU, we discarded it. For each classification method, we explored three number of demand points that corresponded to the number of conservation features explored for prioritizr (3, 6, and 12). By using the same number of demand points per genetic space, we obtained total number of demand points ranging between 51 and 312 (**Table**).

We then explored planning problem where we used all PCA axes to define multidimensional genetic spaces. We placed demand points in these genetic spaces by using the hypervolume method implemented in raptr, which depends on the R package ‘hypervolume’ v. 3.1.1 (Blonder et al 2023). This method first estimates a Gaussian kernel density in the hyperdimensional space defined by the PCA axes and then randomly generates demand points inside a shape that delimits a given quantile of the kernel density. The weight of each demand point is set to the kernel density estimate in that point. We set the quantile to 95% and explored three scenarios for the number of demand points: 20, 40 and 80).

We thus defined a total of 13 planning problems (10 using single PCA axes and 3 using multiple PCA axes) with raptr for both species (**Table**).

In raptr, we assigned a 15% amount representation target to the species distribution and 75% representation targets for each genetic space. We chose the 75% space representation target so that no space had a representation target above the maximum possible target (see comparison of planning solutions – maximum targets). raptr defines planning problems under the minimum set formulation, i.e. finding the selection of PU that maximized the targets met while minimizing conservation costs. We used the Gurobi 9.5.2 optimization tool (REF) and generated a portfolio of 100 solutions within 2% optimality.

## Comparison of planning solutions

We explored how the different definitions of planning problems affected the results of SCP with prioritizr and raptr, focusing on i) selection of PUs ii) meeting conservation targets and iii) conservation cost. For raptr, we also evaluated the maximum target for genetic spaces.

To compare the selections of PUs between planning problems, we first calculated the selection frequency of each PU over the 100 solutions of each problem and used them to calculate pairwise Jaccard distances between problems using function ‘vegdist’ in the R package ‘vegan’ 2.6-4 (Oksanen et al). We tested for significance of the pairwise Jaccard distances through a permutation test: we permuted the 200 solutions of each pair of problems 1000 times, and recalculated the selection frequencies and the Jaccard distance. We then computed the empirical cumulative distribution function of the permuted pairwise Jaccard distances to calculate the p-value of the observed Jaccard distance. Finally, we clustered the planning problems as a function of the observed pairwise Jaccard distances using the complete linkage method in the R function ‘hclust’. We show the results of this clustering with a hierarchical dendrogram.

To evaluate if conservation targets were met, in prioritizr, we calculated the percent amount held for each conservation feature; in raptr, we calculated the percent space held for each genetic space. We did this by crossing each problem with the selections of PUs found in each of them (144 evaluations for prioritizr; 169 evaluations for raptr). In prioritizr, we evaluated whether the solution of a planning problem with conservation features defined with a specific method allowed for meeting the targets of the conservation features defined with the other methods. Similarly, in raptr, we evaluated whether the solution of a planning problem with genetic spaces defined with a specific method allowed for meeting the targets of the genetic spaces defined with the other methods. For each evaluation, we averaged the percent amount (or percent space) held over the conservation features (or genetic spaces) of the problem and the 100 solutions.

Conservation cost was quantified as the total number of PUs selected in a solution.

in RAPTR, we evaluated the maximum target reachable for each genetic space, using the R function ‘maximum.targets’. The maximum target is calculated from equation xxx (s^2) by selecting, for each demand point, the nearest PU. It represents the highest proportion of genetic space reachable by any selection of PUs irrespective of the number of PUs selected (and thus of conservation cost). The maximum target depends on the distribution of demand points in the genetic space and puts an upper limit on the proportion of genetic space that is possible to protect with that configuration of demand points.

## Evaluation and optimal extension of the current set of marine reserves

# Pezzi di discussione

**How many PCA axes?** choice seems to be a compromise between achieving a good representation of genetic diversity (by including more than one PCA axis) and reducing computational complexity (by defining a single space). However, there is no rigorous evaluation of how the inclusion of one or more PCA axes affects the evaluation and planning of PAs. Intuitively, the number of PCA axes retained will be large when PCA axes represent different components of genetic diversity, which depends on a myriad of factors related to the nature of the species and the type of genetic markers. When genetic diversity is related to environmental variables, as in the case of local adaptation, and these variables have different geographic pattern, different PCA axes are likely to express important and different components of variation. In contrast, when a large number of neutral markers distributed throughout the genome (e.g., SNPs) are used, each axis of the PCA is likely to express a small percentage of the total genetic diversity.

**When using single PCA axes as conservation features / genetic spaces: Conservation costs increase with the number of PCA axes included in the analysis only for prioritzr, but not for raptr.** The reason is that prioritizr seeks to attain an amount representation target for all biodiversity features included in the prioritization; as the number of PCA axes increases, so does the number of biodiversity features (because each axis is split into separate layers) and the number of PUs needed to attain the 15% amount target for each of them. Conversely, raptr does not define different features with an amount representation target for each, but only single demand points and an overall amount representation target for the species. As only one planning unit is assigned to each demand point (as per constraint Eq. 3D in Hanson et al 2018), the conservation costs incurred by the genetic targets are lower than that needed to satisfy the overall amount representation target. For example, the most demanding scenario for D. sargus includes 43 PCA axes with 6 demand points each, i.e. 258 demand points in total. This number is well below the overall representation target of 15% of occupied PUs. i.e. 0.15\*2266 = 340 PUs. The total conservation cost in raptr is thus driven by the amount representation target set for the species. In addition, multiple demand points can be satisfied by the same planning unit, further reducing the cost needed to satisfy genetic representation targets.

**Single vs combined PCA axes.** There are potential advantages and disadvantages to considering each axis as a single conservation feature or genetic space. The advantage is that the tasks of partitioning the continuous PU scores into discrete classes or generating demand points are more intuitive when working on one-dimensional values (scores of PUs on one axis of PCA) than when working on multidimensional values (coordinates of PUs in the multidimensional space defined by multiple axes of PCA). The main disadvantage is that, as the number of PCA axes increases, so do the number of conservation features (in prioritizr) and the number of genetic paces (in raptr), and the targets associated to them, potentially creating conservation problems whose solutions are computationally longer to find and involve higher total conservation costs.

**Methods to subdivide continuous values into classes.** Equal interval subdivision involves dividing the range of values into equal-sized intervals. This approach is straightforward, easy to implement, and is useful when the data follows a uniform distribution. However, it may not be suitable when dealing with datasets that have uneven distributions, as it can result in some classes being dominated by data points while others have very few. An extreme case is the creation of classes with no PUs (Figure **S-class-method** B): raptr cannot create a demand point in the empty class because its weight would be zero, since weights are proportional to the number of PUs in a class. In this case, the solution is to remove the empty class, but this results in a lower number of classes than the one set.

Standard deviation subdivision involves dividing the data based on the standard deviation from the mean. It is particularly useful when dealing with datasets that exhibit varying levels of dispersion. However, it assumes a normal distribution of data and may not be suitable for datasets that deviate significantly from normality.

Quantile subdivision, also known as equal percentile or equal probability, involves dividing the data into classes based on the percentiles. Each class contains an equal number of data points. This approach ensures that each class represents the same proportion of the dataset, and is particularly useful when dealing with skewed or unevenly distributed datasets. However, it may result in uneven class intervals in terms of the range of values they cover.

A potential problem affecting equal interval, quantile and standard deviation subdivision in raptr was that, in some cases, the distribution of PU PCA scores was skewed to the boundaries of the class, so that the midpoint of the class was far from the values of the class (Figure **S-class-method** C). Since the midpoint of the class was used as the coordinate for the demand point of the class in raptr, this resulted in a demand point that was far from PU values, and even the nearest PU was relatively far from the demand point. As a consequence, maximum genetic targets (the one achieved when all PUs are selected) were as low as 0.75, when usually could be as high as 0.99 (Figure **S-class-method** D). In these cases, the solution was to set a lower genetic target for that PCA axis, but the problem is that the class does not represent the values well.

The natural breaks or Jenks method aims to create classes by maximizing the differences between the classes while minimizing the differences within each class. It uses an algorithm that iteratively tests different potential breaks in the data range. The goal is to find natural groupings or thresholds where the data exhibits distinct transitions. This method can handle a wide range of data distributions and is suitable for identifying meaningful patterns in the data. However, it can be computationally intensive and may result in unbalanced class sizes. It sometimes returned coincident breaks between classes. For example, Natural breaks with 6 classes resulted in 2 coincident breaks (Figure **S-class-method** A). The solution is to remove one of the coincident breaks, but this results in a lower number of classes than the one set. This affects both prioritizr and raptr

Immagine che contiene testo, diagramma, schermata, Parallelo

Descrizione generata automaticamente

**Figure S-class-method: Problems arising when splitting PCA axis scores into discrete classes.** Panels (A), (B) and (C) shows the frequency distribution of PU scores for a single PCA axis (grey bars) and the breaks between the discrete classes generated by different class methods (red vertical lines). In these examples, all methods were set to generate six discrete classes. In (A), five classes are generated, even if the input number of classes was six, because the last two breaks were coincident at 12.34. In (B), one of the six generated classes (the one bounded between -41 and -30) was empty. In (C), the PU score in the first class were skewed to the lower and higher boundaries of the class: as a consequence, the demand point of the class (red dots) was far from any PU score in that class. Panel (D) shows the consequences of the case shown in (C) on the maximum achievable target for different PCA axes.

**Clustering.** The k-means clustering algorithm is an iterative algorithm used to partition a multivariate dataset into k distinct clusters, aiming to minimize the sum of squared distances between data points and the centroid of their assigned cluster.

Il consenso dice di conservare tutte le popolazioni geneticamente distinte, e al loro interno di conservare una certa porzione di diversità genetica, con cifre proposte che vanno dal 90% al 99%. Ci sono un paio di problemi con questo approccio. Il primo è la difficoltà di identificare popolazioni “distinte” a causa della numerosità dei parametri popolazionistici che interagiscono per determinare la connettività tra le popolazioni, intesa sia dal punto di vista demografico sia dal punto di vista genetico, e della difficoltà di ottenere stime accurate di questi parametri (Lowe and Allendrof 2010; Pasboll 2007; Funk 2012; Barbosa 2018). epIl secondo problema è che E necessario quindi formulare degli indicatori/misure di diversità genetica che descrivano come la diversità genetica è strutturata nello spazio (e non necessariamente tra popolazioni) e dei target di protezione per queste situazioni.