ECOGRAPHY

Research

Integration of presence-only data from several sources: a case study on dolphins' spatial distribution

Sara Martino*, Daniela Silvia Pace*, Stefano Moro, Edoardo Casoli, Daniele Ventura, Alessandro Frachea, Margherita Silvestri, Antonella Arcangeli, Giancarlo Giacomini, Giandomenico Ardizzone and Giovanna Jona Lasinio

S. Martino, Dept of Mathematical Sciences, NTNU, Norway. – D. S. Pace (https://orcid.org/0000-0001-5121-7080) ✉ (danielasilvia.pace@uniroma1.it), S. Moro, E. Casoli, D. Ventura, A. Frachea, M. Silvestri, G. Giacomini and G. Ardizzone, Dept of Environmental Biology, Sapienza Univ. of Rome, Italy. SM also at: Dept of Integrated Marine Ecology, Stazione Zoologica Anton Dohrn, Italy. – A. Arcangeli, ISPRA, Italian Inst. for Environmental Protection and Research, Rome, Italy. – G. Jona Lasinio, Dept of Statistical Sciences, Sapienza Univ. of Rome, Italy.

Ecography
44: 1–12, 2021
doi: 10.1111/ecog.05843

Subject Editor: Miguel Nakamura
Editor-in-Chief: Miguel Araújo
Accepted 19 July 2021

Presence-only data are typical occurrence information used in species distribution modelling. Data may be originated from different sources, and their integration is a challenging exercise in spatial ecology as detection biases are rarely fully considered. We propose a new protocol for presence-only data fusion, where information sources include social media platforms, to investigate several possible solutions to reduce uncertainty in the modelling outputs. As a case study, we use spatial data on two dolphin species with different ecological characteristics and distribution, collected in central Tyrrhenian through traditional research campaigns and derived from a careful selection of social media images and videos. We built a spatial log-Gaussian cox process that incorporates different detection functions and thinning for each data source. To finalize the model in a Bayesian framework, we specified priors for all model parameters. We used slightly informative priors to avoid identifiability issues when estimating both the animal intensity and the observation process. We compared different types of detection function and accessibility explanations. We showed how the detection function's variation affects ecological findings on two species representatives for different habitats and with different spatial distribution. Our findings allow for a sound understanding of the species distribution in the study area, confirming the proposed approach's appropriateness. Besides, the straightforward implementation in the R software, and the provision of examples' code with simulated data, consistently facilitate broader applicability of the method and allow for further validations. The proposed approach is widely functional and can be considered with different species and ecological contexts.

Keywords: cetacean, data fusion, dolphins, Mediterranean Sea, presence-only data, point processes

Introduction

Progress of ecological science is more and more reliant on combining data from diverse sources (Fletcher et al. 2019). This approach can increase the comprehension

*Equally contributed.

NORDIC SOCIETY OIKOS
www.ecography.org

of ecological processes for both research and conservation purposes (Pace et al. 2019). Data availability to model species distribution, for example, is rapidly expanding thanks to the fast development of new technologies (Soranno and Schimel 2014), the growth of citizen science initiatives (Sicacha-Parada et al. 2020, Matutini et al. 2021) and the opportunity of exploiting huge information harvested from social media platforms (Mikula and Tryjanowski 2016, Pace et al. 2019). The latter data types can be intrinsically challenging to merge in with existing, valued and validated data collected via standard research protocols. Yet, if that can be achieved, they can offer enrichment of existing data to generate powerful insights and even reduce the costs of collecting data conventionally (Buchanan and Bryman 2018). Nevertheless, heterogeneous data are complex to manage as they are polymorphic in nature and affected by numerous forms of bias and limitations (Isaac and Pocock 2015). Information on species occurrence collected at sea by sea-users, for example, is characterised by a different spatiotemporal distribution of effort, which can be biased toward easily accessible habitat and times with better weather, or known areas of use (Corkeron et al. 2011, Sicacha-Parada et al. 2020). Hence, a simple data pooling (Fletcher et al. 2019) with data gathered under conventional research methodologies is not enough to reliably model the presence of a species considering different explanatory variables both environmental and anthropogenic and to define its distribution over multiple spatial and temporal scales.

Integrated distribution modelling (IDM), i.e. the practice of fitting species distribution models with more than one observation practice (Isaac et al. 2020), is a new approach to combine different datasets, preserving the strengths of each and adjusting, at least to some extent, their limitations. IDM sets a spatial or spatio-temporal latent state, here statistically defined as a point process, of the sites where the animals were sighted, described by a series of covariates shared by different datasets. Multiple observation sub-models can be estimated from them, each describing a part of the latent state.

Coping with several challenges, we propose a novel path to combine different data sources to provide cohesive summaries of species' potential and realised distribution (Isaac et al. 2020). First, as the available information is presence-only data, we opt for point process as the most natural solution (Miller et al. 2019). Second, as several sources of bias are potentially present in the datasets, we propose models based on a location-dependent thinning of a Poisson process to reduce these biases (Dorazio 2014, and references therein); however, the parameters of these models are not fully identifiable unless the covariates of abundance are distinct and linearly independent of the covariates of detectability (Dorazio 2014). In Yuan et al. (2017), a flexible stochastic partial differential equation (SPDE) model describes the spatial structure that is not accounted for by explanatory variables, and estimation is carried on using integrated nested Laplace approximation (INLA) in a Bayesian inference framework. The latter allows simultaneous fitting of detection and density models and permits prediction at an arbitrarily fine scale. Very recently Sicacha-Parada et al. (2020) adopt a similar approach using citizen science data on moose *Alces alces* occurrence in Norway, accounting for the geographical bias (oversampling of 'accessible' locations). For marine observations, the boat's size, the distance from the coast, policy regulations and weather conditions are just some of the factors that can affect the accessibility of an area. We propose a new protocol for presence-only data fusion, where information sources include social media. We investigate several possible solutions and compare different types of detection function and accessibility explanations. We use the IDM approach on sighting data derived from different data sources (research, monitoring and social media) to predict the distribution of two dolphin species in the central Mediterranean Sea. The study of spatial distribution patterns of dolphin species is incredibly puzzling. They spend much time under the water surface (Redfern et al. 2006), and a lot of visual/acoustic effort for scientists is needed to assess their presence in a specific habitat (Breen et al. 2017, Redfern et al. 2017). We show how variation in the detection function affects ecological findings on two dolphin species with different spatial distribution.

The proposed approach is entirely broad and the selected species are representatives for different habitats. Hence they constitute a good benchmark for the entire proposal. We provide R functions and example code to replicate our work in the online Supporting information (<https://github.com/smar-git/SM-data-merging>).

Material and methods

Study species

Two dolphin species were selected for this study, the bottlenose *Tursiops truncatus* and the striped dolphins *Stenella coeruleoalba*, both widely distributed throughout the Mediterranean Sea. The bottlenose dolphin is reported predominantly coastal or inshore (Bearzi et al. 2012), but its habitat changes depending on the region: it can inhabit shallow waters (less than 50 m) close to the coast and at the mouths of the rivers (Triossi et al. 2013, Pace et al. 2019), around archipelagos or islands (Pace et al. 2012, 2019, Pulcini et al. 2014), and in waters above the continental shelf and slope (Azzellino et al. 2008); less frequent, but still present, in deeper waters and pelagic areas. Bottlenose dolphins feed a wide range of demersal and coastal prey and can forage opportunistically behind trawling vessels (Pace et al. 2012). The striped dolphin is considered pelagic in the Mediterranean Sea, showing a general preference for highly productive, open waters beyond the continental shelf (Aguilar and Gaspari 2012). Although the species is the most abundant cetacean in the Mediterranean, it is not found at uniform densities. The striped dolphin diet is mainly composed by pelagic or bathypelagic schooling-nictemeral fish, squids and even crustaceans (Meissner et al. 2012). There are not exact estimations of the number of bottlenose and striped dolphins living in the Mediterranean Sea. The poor understanding of the status of a population, together with the suspected decline in numbers (both species

are listed under the status vulnerable in the IUCN Red List as their populations have been decreasing during the last decades), emphasize the importance of integrating all available information (Pace et al. 2014, 2021b).

Study area

The study area covers about 39 000 km^2, and is located in the central Tyrrhenian Sea (Italy) (Fig. 1); it is characterized by different environmental features (e.g. bathymetries), structures (e.g. seamounts) and types of habitats (Pace et al. 2019, 2021a). Several rivers flow in this region, including the Tiber, and the simultaneous presence of both fresh and salt waters, as well as the geomorphological action of sedimentation and erosion, generate different ecological gradients, making the coastal area highly productive and rich in biodiversity (Ventura et al. 2015, Ardizzone et al. 2018, Casoli et al. 2019). The study area also includes five islands (Giglio and Giannutri at north; Ponza, Ventotene and Santo Stefano at south) and several commercial/touristic harbours generating high-levels of maritime traffic by different vessels. The region hosts seven of the eight cetacean species regularly found in the Mediterranean, with a major presence of bottlenose and striped dolphins (Pace et al. 2019, 2021a).

Data sources and attributes

Dolphin data cover a period of 13 years (2007–2019). Records are from three sources: a) conventional research protocols from motor and sailing boats (non-systematic haphazard, sensu Corkeron et al. 2011) (labelled UNIRM) (Pace et al. 2019); b) standardized monitoring protocols from platforms of opportunity within the project FLT Mediterranean Monitoring Network (labelled FERRY) (ISPRA 2016, Arcangeli et al. 2019, Pace et al. 2019); c) social media reports (Facebook and YouTube) by sea-users (Pace et al. 2019) (labelled SM). Data collection protocols and selection procedures are provided in Pace et al. (2019). As the SM dataset included also details on other cetacean species than the two here investigated (Fig. 1b), we used this information as a proxy to infer boat densities potentially able to record the animals' presence.

These three sources accounted for 283 records of striped dolphin (about 50% from SM) and 579 of bottlenose dolphin (about 80% from SM). The major contribution by SM justified the need for a careful choice of the related model's elements.

We used distance from the coast (i.e. the euclidean distance between a sighting point and the shoreline), depth, slope, temperature and primary productivity as covariates. These are commonly selected in cetacean distribution studies as they may represent good proxies for species' ecological needs (Chavez-Rosales et al. 2019, Stephenson et al. 2020). Temperature and primary productivity were retrieved from COPERNICUS platform <https://marine.copernicus.eu/>. Depth data were downloaded from GEBCO (General bathymetric Chart of the Ocean – <www.gebco.net/>). Slope was computed from depth data through the `terrain()` function in R. Details of the retrieved datasets and covariates handling procedures are reported in the Supporting information.

Modelling approach

Our aim was to integrate data from three main sources. Two are typical approaches adopted in research surveys. We consider the adaptive sampling procedure used by Sapienza University of Rome (UNIRM) (see for instance Dawson et al. 2008, Lennert-Cody et al. 2018, and references therein) and the very well known distance sampling (Buckland et al. 2001) adopted by ISPRA (FERRY) (ISPRA 2016), together with the Social Media (SM) extracted data (below and Pace et al. 2019, for detailed description of the data). We aimed at representing and managing possible detection bias in each dataset adopting a point processes modelling approach. The Supporting information illustrates and summarises the workflow used for building the model.

We followed Yuan et al. (2017) and Sicacha-Parada et al. (2020), expanding their approaches by building a spatial log-Gaussian cox process (LGCP) (Illian 2019) that incorporates different detection functions and thinning for each data source. We assumed that sighting patterns, i.e. locations of dolphin groups in space ($s \in \mathcal{S} \subset \Re^2$) and time ($t \in \mathcal{T}$), are properly described by a point process whose intensity function $\lambda(s,t)$ is additive on the log-scale:

$$\log\left(\lambda(s,t)\right) = \mathbf{X}^T(s,t)\beta + f(\mathbf{z}) + \omega(s) \tag{1}$$

Here $\mathbf{X}(s,t)$ is a set of covariates detected at location s and time t with linear effect β to be estimated. $f(\mathbf{z})$ is a smooth effect (that may be present or not) of some geo-referenced covariates $\mathbf{z}$. A common prior for $f(\mathbf{z})$ is a random walk (RW) model of order 1 (Rue and Held 2005). Finally, $\omega(s)$ is a zero-mean Gaussian process describing the residual spatial variation. As in Yuan et al. (2017) we adopted a Matèrn covariance of order 1 with range ρ and standard deviation σ. Although it would have been, in theory, possible to consider $\omega(s)$ a complex spatio-temporal model (Yuan et al. 2017), the limited number of sightings each year did not provide enough information in practice. Therefore we chose to run $\omega(s)$ a pure spatial model.

We assumed that the above process was observed in three different ways, conditionally independent given $\lambda(s,t)$. Thus, three observed intensities were defined:

$$\lambda^*(s,t) = T_j g_j(s)\lambda(s,t), \qquad j = 1,2,3 \tag{2}$$

where T_j is a time scaling factor and $g_j(s)$ is the detection function (with values between 0 and 1) which determines the thinning of the original process. The form of the detection function depends on the type of observational process. For adaptive sampling (UNIRM)

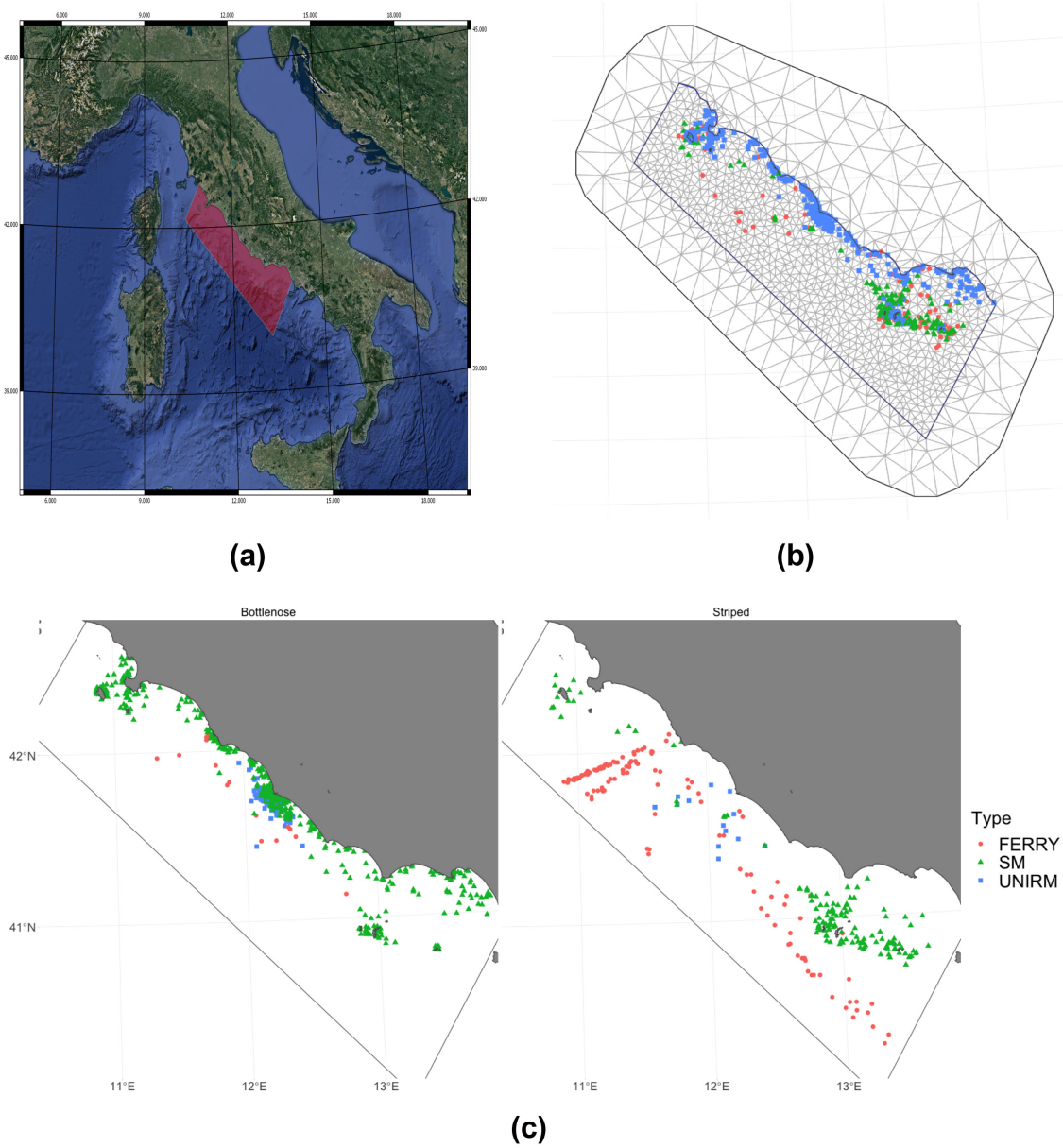

Figure 1. (a) Study area. (b) Study area and SM records for striped (green triangles), bottlenose (blue squares) dolphins and other cetacean species (red dots) superimposed to the mesh chosen for models estimation. (c) Dolphins encounters' locations by source (observation processes): SM (green), FERRY (red) and UNIRM (blue).

$$g_1(s) = \begin{cases} 1 & d_1(s) \le K \\ 0 & d_1(s) > K \end{cases} \tag{3}$$

where $d_1(s)$ is the distance (km) between point s and the position of the boat when the groups were sighted. K was defined as the maximum distance measured between the location of the first visual sight of a dolphin group by researchers (equipped with 7 × 50 and 10 × 50 binoculars) on the boat and the effective location of the group under optimal survey conditions (i.e. sea state ≤ 1 Douglas, wind force ≤ 1 Beaufort, no rain, no fog, no clouds). This measurement was possible because, upon sighting dolphins, researchers marked the GPS point where the animals were first located, the survey effort was suspended and the vessel departed from its route to approach the group to a suitable distance (10–30 m) to correctly identify the species, estimate group size and composition. K was set to 4 km, assuming that researchers can spot animals closer than K.

For the distance sampling (FERRY) data, we used the classical half normal detection function (Thompson and Ramsey 1987) defined as

$$g_2(s) = \exp\left(-\frac{d_2(s)^2}{2\,\xi_2^2}\right) \tag{4}$$

where, $d_2(s)$ is the perpendicular distance (km) to the ferry track and ξ_2 is a scale parameter.

For the SM dataset, the definition of the detection function was carefully considered for biases. Records in this dataset are affected by large uncertainty, as observations are generally a) skewed towards more accessible areas (Monsarrat et al. 2019, Sicacha-Parada et al. 2020) and b) collected from small leisure boats that are difficult to track in a systematic way. To better define 'more accessible' and consider the distribution of the small boats we explored three different possibilities.

First, we reasonably assumed that locations closer to the coast are more accessible to sea-users with small boats. Thus, following Sicacha-Parada et al. (2020), the detection function, labelled as 'detection coastline', was defined as:

$$g_{3,1}(s) = \exp\left(-\frac{d_{3,1}^2(s)}{2\,\xi_{3,1}^2}\right) \tag{5}$$

where $d_{3,1}(s)$ is the Euclidean distance from the coast (Fig. 2a) and $\xi_{3,1}$ a scaling parameter. However, the distance from the coast may not provide an accurate representation of the small boats' density in a given area: locations close to harbours and holiday destinations (e.g. islands) are generally more crowded than other sites at the same distance from the coastline.

To obtain information on the boats density in the study area, we used data from EMODnet (European Marine Observation and Data Network; Martín Míguez et al. 2019), a free-usage platform of vessel density data derived from boats using AIS (automatic identification system, mandatory above 15 m length). The database has a spatial resolution of 11 km and covers 2017–2019 period. We selected two vessels categories (sailboats and pleasure crafts) from the 11 listed, and applied a kernel estimator to ensure a smoothed density surface. The resulting log-density surface (Fig. 2b) was labelled as vessel log-density surface. As expected, higher vessel log-densities were identified near the principal harbours and the islands. Our second detection function for SM data, denoted 'detection Emodnet', was defined as

$$g_{3,2}(s) = \Phi\left(\frac{d_{3,2}(s)}{\xi_{3,2}} - \mu_{3,2}\right) \tag{6}$$

where $d_{3,2}(s)$ is the vessel log-density, and Φ is the normal cumulative distribution function (cdf) with $\mu_{3,2}$ and $\xi_{3,2}$ as location and scale parameters, respectively. The normal cdf was selected as we required the detection function to be close to 1 when the vessel log-density is high, and close or equal to zero when it is small (or null). EMODnet information

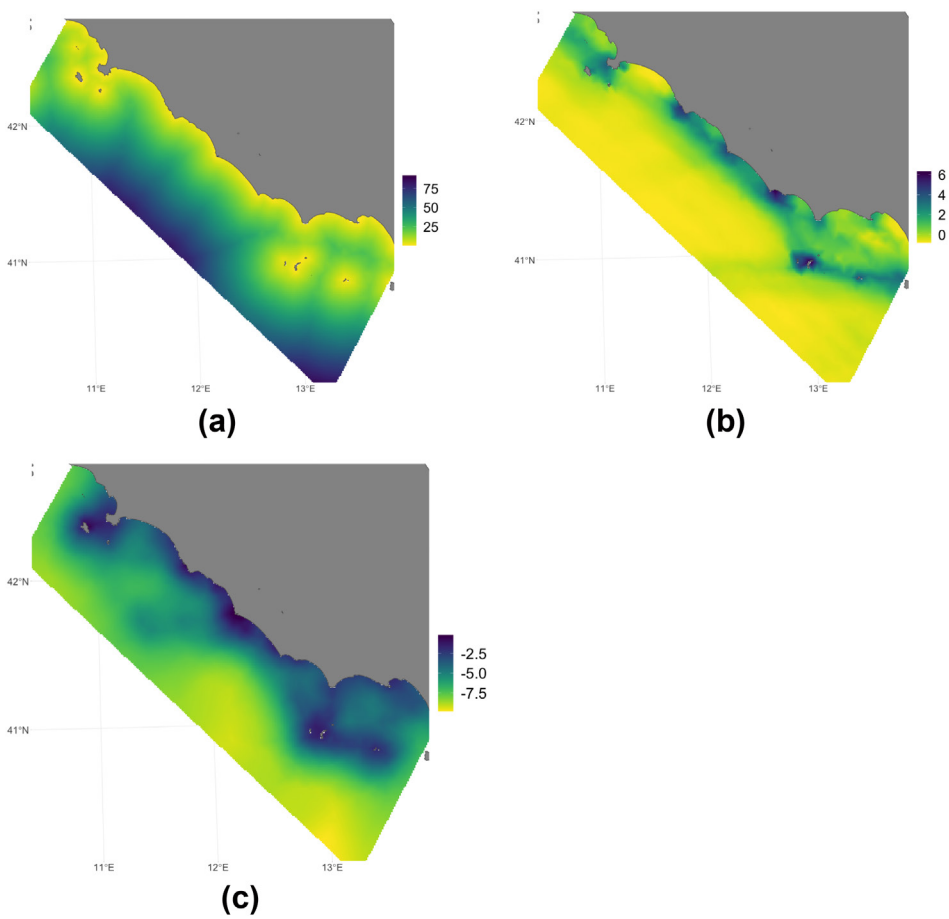

Figure 2. (a) Distance from the coastline (km), (b) log vessels density, (c) estimated log intensity from observations of all species.

accounted for a limited time frame compared to our study and for larger vessels than the ones generally reporting observation records in SM platforms (small recreational boats moving near the coastline). We therefore introduces a third detection function.

We used the entire SM dataset of 581 records (125 striped and 334 bottlenose dolphins, and 122 other cetacean species) to estimate the observation process intensity. We considered the spatial pattern of such observations as a proxy for the small boat density process if we disregard the species. A similar approach was used in occupancy models context, where non-detection records were constructed from sightings of other benchmark species (Kery et al. 2010, Dennis et al. 2017). We applied a spatial LGCP to estimate the (log) intensity of the process. Details of the estimation process can be found in the Supporting information. Figure 2c shows the resulting estimated log-intensity used as input for the detection function, labelled 'detection animals':

$$g_{3,3}(s) = \Phi\left(\frac{d_{3,3}(s)}{\xi_{3,3}} - \mu_{3,3}\right) \tag{7}$$

where $d_{3,3}(s)$ is the estimated log-intensity at point s while Φ, $\mu_{3,3}$ and $\xi_{3,3}$ are defined as in (6).

Eventually, another potential bias affecting the observation processes is the different time (days) spent at sea by each data source. To account for this bias as well, we introduced the t_j parameter in expression (1). T_j is known for both the FERRY and the UNIRM data (311 and 73 days at sea respectively) and undetermined for SM data. We know that SM observations were collected by leisure boats all over the year, with a major number of sightings reported in spring–summer. Thus, we ran estimations with $T_3 = 160$, 200, 365 days, without sensible changes, and selected $T_3 = 360$.

Table 1. Comparison criteria for the four fitted models for both striped (a) and bottlenose (b) dolphins.

Model	DIC	WAIC	MLIK	LMPL
(a) *Stenella coeruleoalba*				
Constant detection	4078.53	4129.16	−2111.81	−2098.13
Detection coastline (Eq. 5)	3895.52	3933.01	−2008.44	−1988.77
Detection emodnet (Eq. 6)	3840.93	3889.51	−2019.20	−1969.47
Detection animals (Eq. 7)	3789.38	3810.08	−1942.07	−1922.56
(b) *Tursiops truncatus*				
Constant detection	4639.94	4874.47	−2375.33	−2568.07
Detection coastline (Eq. 5)	4555.23	4797.14	−2337.60	−2726.91
Detection emodnet (Eq. 6)	4552.61	4810.37	−2344.68	−2658.98
Detection animals (Eq. 7)	4485.78	4624.58	−2281.27	−2351.19

Priors specification

To finalize the model in a Bayesian framework, we needed to specify priors for all model parameters. To avoid identifiability issues when estimating both the animal intensity and the observation process, we used slightly informative priors. For the parameters in the spatial field $\omega(s)$ in (1) we used PC priors (Fuglstad et al. 2019) setting $P(\rho < 150) = 0.5$ and $P(\sigma > 2) = 0.01$, thus we considered a standard deviation above 2 and a range of 150 km likely. We assigned β in (1) and the locations parameters $\mu_{(3,2)}$ and $\mu_{(3,3)}$ Gaussian prior precision 0.01 and means 0, 3 and −5 respectively. Finally, for the scale parameters in (4–6), let $\xi = F_{\alpha}^{-1}(\Phi(\theta))$ where $F^{-1}(\cdot)$ is the inverse exponential cdf with rate α and Φ a normal cdf. This corresponds to attributing an exponential prior to ξ. We assigned θ a standard normal prior. The parameter α is set to 1/20 in (5), and 1 in all other cases. The difference in rate was due to the different scale of the three inputs for the detection function (Fig. 2). The Supporting information illustrates the effect of our prior choice on the detection functions.

Inference and computational approach

The traditional way of fitting point processes is by gridding the space and modelling the intensity on a discrete number of cells. This implies that observations' locations are also approximated. We followed instead the approach introduced in Simpson et al. (2016) and applied in Yuan et al. (2017) and Sicacha-Parada et al. (2020). Such an approach allowed us to use the true sighting locations, thus avoiding loss of information. Besides, the Gaussian field's SPDE representation has several computational advantages (Lindgren et al. 2011). To build a spatial model using the SPDE approach, we used the mesh shown in Fig. 1b.

For computational efficiency, we used INLA (Rue et al. 2009). INLA allowes also to easily combine the three observation model in (2) to form the likelihood. Our model does not directly fall under the latent Gaussian model framework for the INLA estimation software because the parameters in the detection functions in (4–6) do not enter the model in a log-linear way. We used therefore the methodology introduced in Yuan et al. (2017) and implemented in the `inlabru` R package (Bachl et al. 2019) that allows fitting models with some non-linear elements. This is done by linearizing the model via Taylor approximation and using a line search to optimize the linearization point.

Model evaluation was carried out using goodness of fit measures as in Sicacha-Parada et al. (2020), through the deviance information criterion (DIC), Watanabe–AiKaike information criterion (WAIC), marginal likelihood (MLIK) and the logarithm of the pseudo marginal likelihood (LMPL). As a benchmark for the SM detection function choice, we used a constant detection function $g(s) = 1, \forall s$, that is equivalent to not include any thinning for the SM data.

Results

The distribution of the dolphins encounters in the study area is shown in Fig. 1c. Environmental covariates selection was finalized considering several combinations of covariates and detection functions. Two different models have been selected, one for each species

S. coeruleoalba (striped dolphin)
- Depth: categorized as (< 100, 100–200, 200–1000, > 1000 m)
- Slope: non parametric with a prior Random walk of order 1
- Distance from the coast: linear term

T. truncatus (bottlenose dolphin)
- Depth: linear term
- Slope: linear term
- Distance from the coast: linear term

There was no evidence that the sightings intensity was affected by the spatio-temporal covariates, therefore our final models are reduced to purely spatial ones.

The evaluation of SM detection functions was based on model's goodness of fit measures, DIC, WAIC, MLIK and the LMPL, it is reported in Table 1. The selected best performing detection function for all criteria and species is (7) (labelled as 'intensity'). This choice affected model's terms estimate. For striped dolphin model with varying detection functions (Supporting information), the effects of categorized depth were fairly in agreement with the species distribution ranges: it is generally not found in very shallow waters (negative effects), observed at 100–200 m depth, and more often encountered at depths over 200 m.

The effect of the detection function was found in the reduction of uncertainty in the estimates, which is reflected in the smaller size of the credibility intervals (7). Slope showed a significant reduction effect in the encounters where it is steeper. No significant difference was found among the smooth effects with varying detection (overlapping 95% confidence band, not shown). The effect of the Distance from the coast was not significant, and the intercept was larger for detection functions (6) and (7), with the latter showing less uncertainty than the first.

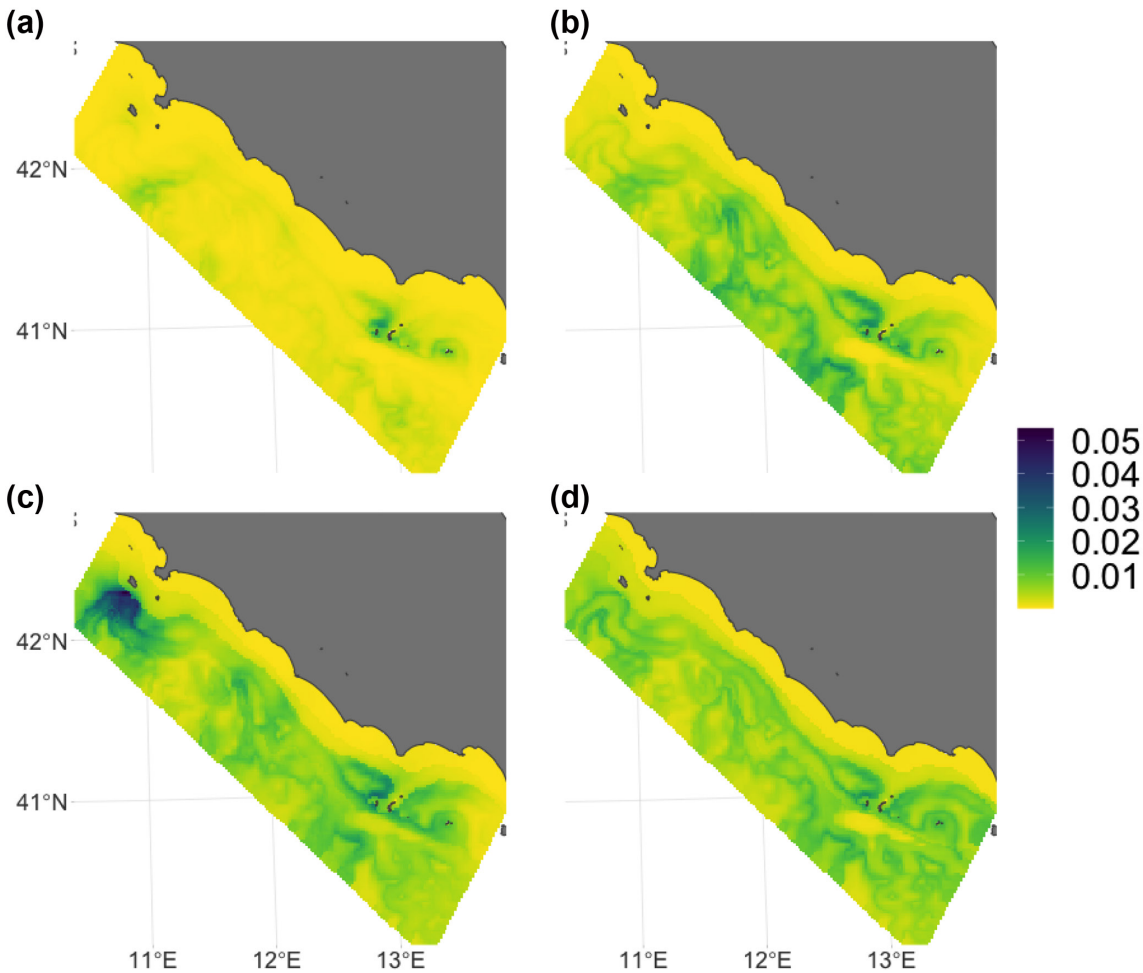

Figure 3. Estimated posterior median for the intensity of striped dolphins using different detection functions for SM data. (a) constant detection, (b) detection coastline (Eq. 5), (c) detection Emodnet (Eq. 6), (d) detection animals (Eq. 7).

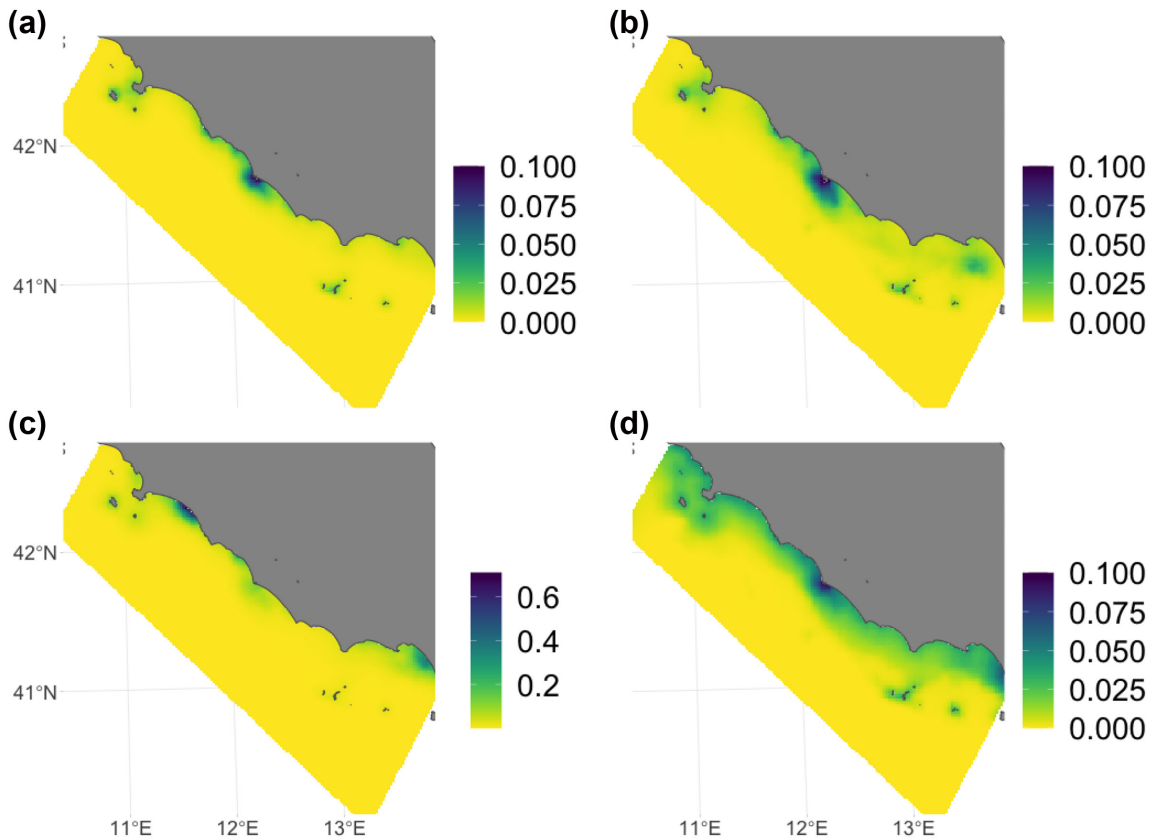

Figure 4. Estimated posterior median for the intensity of bottlenose dolphin using different detection functions for SM data. (a) constant detection, (b) detection coastline (Eq. 5), (c) detection vessels (Eq. 6), (c) detection animals (Eq. 7). Note that the scale in (c) is different from the other three figures.

For bottlenose dolphin model with varying detection functions (Supporting information), both depth and distance from the coast had negative effects on sightings (deeper waters and increasing distance from the coast mean less encounters), with no-significant difference among detection functions. Again, detection (7) induced narrower 95% credible intervals.

Estimates of detection functions parameters for both species are reported in the Supporting information.

As a measure of relative uncertainty for the predicted intensity $\lambda(s)$ we used the relative width of the 50% posterior credible interval (RWPCI) as proposed in Yuan et al. (2017). This measure is defined as the interquartile range divided by the median

$$\mathrm{RWPCI} = (Q_3 - Q_1) / Q_2 \tag{8}$$

The intensity surfaces estimated for the striped and bottlenose dolphins are shown in Fig. 3, 4, respectively; associated RWPCIs (8) are mapped in the Supporting information. The intensity surface for both species changed consistently with the different detection function adopted for SM. For example, the vessel-based detection (6) (EMODnet-log density surface) induced some artifacts for the striped dolphin (Fig. 3c), and in general seems to over-estimate the dolphins encounter probability. This is evident, for example, around the Giglio Island, where detection highlights hotspots for both striped and bottlenose (Fig. 4c) dolphins possibly induced by the presence of few vessels and several encounters. The detection based on distance from the coast (5) and the constant detection, under-estimate the same probability for the striped dolphin and create some artifacts as well. A relevant feature of the detection function (7) is that it allowed for a consistent reduction in the uncertainty associated with the estimated intensity surface.

Figure 5 describes the estimated probability of the average number of sightings in the area over 13-year study period. These distributions represent the potential encounters if the entire area would be surveyed. (d) corresponds to the chosen detection, as it best represents the studied phenomena. Striped dolphin is considered the most abundant and common cetacean in the Mediterranean (Aguilar and Gaspari 2012), but seems to be less represented in the study area than bottlenose dolphins (283 records of striped versus 579 of bottlenose dolphins). Although this may introduce a large uncertainty on the estimates, it is still possible to appropriately capture the species spatial distribution. In panel (c) an over-estimation of the bottlenose dolphin encounters given by the vessel detection seems evident, and in panel (b) the coastline detection apparently induces a distribution of potential sightings only driven by the data.

Discussion

This study demonstrates that methods of spatial data integration able to carefully consider and minimize datasets' biases can be efficiently used to predict species' distribution. Results here obtained may be broadly applicable to other species that require an improvement of spatial knowledge for their conservation and management.

Dorazio (2014) pointed out that several statistical models have been proposed to analyse presence-only data, but they have largely ignored the effects of imperfect detectability and survey bias. The same author showed that proper modelling choices could reduce the bias in SDM estimates induced by these types of errors. Here we do more than just correct for detectability issues; we allow multiple sources of information to be integrated. We defined and estimated source-specific detection functions considering the nature of the data, i.e. presence-only, and the different observation processes, offering a more precise picture of the distribution of two dolphin species in the central Mediterranean. The output is consistent with the ecology of these species, highly supporting a thoughtful usage of spatial data extracted from social media platforms and introducing a novel way to model observation biases. In analysing different detection functions, we optimise distribution models for each species. That is very attractive considering the importance of defining suitable habitats for vulnerable or endangered cetaceans exposed to anthropogenic disturbance or threats, particularly in coastal areas (Pace et al. 2018).

The point process approach allows us to reliably estimate the observation intensity surface. The analysis of intensity surfaces in Fig. 3, 4, gives important insights on the relevance of the detection function in observation intensity estimation. The artefacts around the Tiber river estuary (central part of the area) for the bottlenose dolphin and close to the Giglio island (northern portion of the study area) for the striped dolphin are solved by detection (7). Again, with the same detection function's choice, analysing in the Supporting information, we can observe the reduction of intensity estimates' variability (and hence uncertainty). The proposed 'best' choice is very general and can be adopted whenever social media data are available.

The two species were also studied in Pace et al. (2019) using a presence-only data approach based on MaxEnt (Phillips et al. 2006). While results related to the bottlenose dolphin analysis were ecologically sound and coherent, striped dolphins analysis was unfeasible in that framework, given the relevant number of near-to-the-coast observation by sea-users. In particular, the depth around the Pontine islands rapidly increases with the distance from the coast, playing a misleading role in the MaxEnt modelling approach. The proposed methodology, instead, is fully able of capturing both species behaviour, thus addressing the complex task of finding targeted techniques weighting species' diversity.

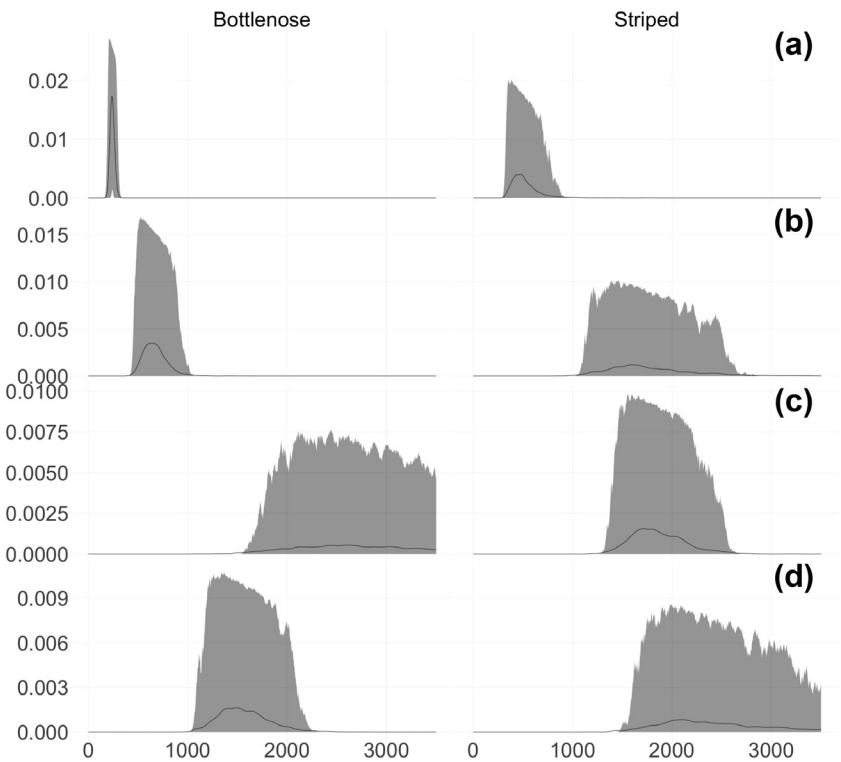

Figure 5. Estimated probability of the average number of sightings in the area over 13-year study period, in the entire study area for bottlenose (left) and striped (right) dolphins for the four fitted models: (a) constant detection, (b) detection (5), (c) detection (6) and (d) detection (7). The grey band indicate 95% credible intervals.

Some limitations are intrinsic to the proposed approach. On the one hand, spatial estimation does not distinguish between land and sea. That implies the use of post-processing to cut the estimated intensity surface. On the other, each analysed detection function is not very flexible. Eventually, the information used to model the observation effort in the SM data can be further improved. Hence, further investigations will be carried out to:

- Develop spatially non-stationary modelling approaches where a barrier can be added at the coastline as in Bakka et al. (2019).
- Develop flexible detection functions.
- Explore the use of satellite data to estimate the density of small boats in the study area (Santamaria et al. 2017).
- Explore the use of biological driving variables (e.g. prey biomass) as predictors.

The implementation of these tasks and the improvement of the models capabilities may further develop a fast-growing research approach and provide innovative insights in marine top-predators distribution patterns. The multiplicity of issues confronting these marine species requires collaborative efforts at all levels to share and merge resources, data and expertise efficiently (Pace et al. 2018, Vella et al. 2021).

Acknowledgments – The authors would like to thank the two referees and the associate editor that provided very valuable comments to improve a previous version of the paper.
Funding – This work was partially supported by project 'Joint Cetacean Database and Mapping (JCDM) in Italian waters: a tool for knowledge and conservation', Sapienza University of Rome (no. RM1201729F23D51B).

Author contributions

Sara Martino: Conceptualization (equal); Data curation (lead); Formal analysis (lead); Methodology (equal); Software (equal); Validation (equal); Writing – original draft (equal); Writing – review and editing (equal). **Daniela Silva Pace**: Conceptualization (lead); Data curation (lead); Investigation (lead); Methodology (equal); Resources (equal); Validation (equal); Writing – original draft (equal); Writing – review and editing (lead). **Stefano Moro**: Data curation (equal); Formal analysis (supporting); Methodology (equal); Software (equal); Validation (supporting); Writing – review and editing (supporting). **Edoardo Casoli**: Data curation (supporting); Formal analysis (supporting); Investigation (supporting); Validation (equal); Visualization (equal); Writing – original draft (supporting); Writing – review and editing (supporting). **Daniele Ventura**: Data curation (supporting); Formal analysis (supporting); Investigation (supporting); Validation (equal); Visualization (equal); Writing – original draft (supporting); Writing – review and editing (supporting). **Alessandro Frachea**: Data curation (equal); Formal analysis (equal); Investigation (supporting); Methodology (supporting); Writing – original draft (supporting). **Margherita Silvestri**: Data curation (supporting); Formal analysis (supporting); Investigation (equal); Methodology (supporting); Writing – original draft (supporting). **Antonella Arcangeli**: Data curation (equal); Investigation (equal); Methodology (equal); Resources (equal); Validation (equal). **Giancarlo Giacomini**: Data curation (supporting); Investigation (equal); Methodology (supporting); Resources (equal); Writing – review and editing (supporting). **Giandomenico Ardizzone**: Conceptualization (equal); Funding acquisition (equal); Resources (equal); Supervision (equal); Validation (equal); Writing – review and editing (supporting). **Giovanna Jona Lasinio**: Conceptualization (lead); Data curation (equal); Formal analysis (equal); Methodology (equal); Software (equal); Supervision (lead); Validation (lead); Writing – original draft (equal); Writing – review and editing (lead).

Data availability statement

Data and tutorials for elaboration are available from: <https://github.com/smar-git/SM-data-merging>. (Martino et al. 2021).

References

Aguilar, A. and Gaspari, S. 2012. *Stenella coeruleoalba*. The IUCN red list of threatened species 2012: e.t20731a2773889. – <https://www.iucnredlist.org/species/20731/50374282>.

Arcangeli, A. et al. 2019. Fixed line transect mediterranean monitoring network (flt med net), an international collaboration for long term monitoring of macro-mega fauna and main threats fixed line transect mediterranean monitoring network. – Biol. Mar. Mediterr. 26: 400–401.

Ardizzone, G. et al. 2018. Atlante degli Habitat dei Fondali Marini del Lazio. – Sapienza Univ. Editrice, Rome, Italy.

Azzellino, A. et al. 2008. Habitat use and preferences of cetaceans along the continental slope and the adjacent pelagic waters in the western Ligurian sea. – Deep Sea Res. Part 1 55: 296–323.

Bachl, F. E. et al. 2019. Inlabru: an R package for bayesian spatial modelling from ecological survey data. – Methods Ecol. Evol. 10: 760–766.

Bakka, H. et al. 2019. Non-stationary Gaussian models with physical barriers. – Spat. Stat. 29: 268–288.

Bearzi, G. et al. 2012. *Tursiops truncatus*. The IUCN red list of threatened species 2012: e.t22563a2782611. – <https://www.iucnredlist.org/species/22563/156932432>.

Breen, P. et al. 2017. Where is the risk? Integrating a spatial distribution model and a risk assessment to identify areas of cetacean interaction with fisheries in the northeast Atlantic. – Ocean Coast. Manage. 136: 148–155.

Buchanan, D. and Bryman, A. 2018. Unconventional methodology in organization and management research chapter not another survey: the value of unconventional methods. – Oxford Univ. Press.

Buckland, S. et al. 2001. Introduction to distance sampling: estimating abundance of biological populations. – Oxford Univ. Press.

Casoli, E. et al. 2019. Comparative analysis of mollusc assemblages from different hard bottom habitats in the central Tyrrhenian sea. – Diversity 11: 74.

Chavez-Rosales, S. et al. 2019. Environmental predictors of habitat suitability and occurrence of cetaceans in the western North Atlantic ocean. – Sci. Rep. 9: 1–11.

Corkeron, P. et al. 2011. Spatial models of sparse data to inform cetacean conservation planning: an example from Oman. – Endang. Species Res. 15: 39–52.

Dawson, S. et al. 2008. Design and field methods for sighting surveys of cetaceans in coastal and riverine habitats. – Mammal Rev. 38: 19–49.
Dennis, E. B. et al. 2017. Efficient occupancy model-fitting for extensive citizen-science data. – PLoS One 12: e0174433.
Dorazio, R. M. 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. – Global Ecol. Biogeogr. 23: 1472–1484.
Fletcher Jr, R. J. et al. 2019. A practical guide for combining data to model species distributions. – Ecology 100: e02710.
Fuglstad, G. et al. 2019. Constructing priors that penalize the complexity of gaussian random fields. – J. Am. Stat. Assoc. 114: 445–452.
Illian, J. B. 2019. Spatial and spatio-temporal point processes in ecological applications. – In: Gelfand, A. E. et al. (eds), Handbook of environmental and ecological statistics. Chapman and Hall – CRC Press Taylor & Francis Group, pp. 97–132.
Isaac, N. et al. 2020. Data integration for large-scale models of species distributions. – Trends Ecol. Evol. 35: 56–67.
Isaac, N. J. B. and Pocock, M. O. 2015. Bias and information in biological records. – Biol. J. Linn. Soc. 115: 522–531.
ISPRA 2016. Fixed line transect monitoring using ferries as platform of observation for marine mega and macro fauna and main threats. Monitoring protocol for cetaceans and sea turtles. – ISPRA Agreement – Technical annex 1 ISPRA, pp. 19.
Kery, M. et al. 2010. Site-occupancy distribution modeling to correct population-trend estimates derived from opportunistic observations. – Conserv. Biol. 24: 1388–1397.
Lennert-Cody, C. et al. 2018. Review of potential line-transect methodologies for estimating abundance of dolphin stocks in the eastern tropical pacific. – J. Cetacean Res. Manage. 19: 9–21.
Lindgren, F. et al. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. – J. R. Stat. Soc. B Stat. Methodol. 73: 423–498.
Martín Míguez, B. et al. 2019. The European marine observation and data network (EMODNET): visions and roles of the gateway to marine data in Europe. – Front. Mar. Sci. 6: 1–24.
Martino, S. et al. 2021. Data from: Integration of presence-only data from several sources: a case study on dolphins' spatial distribution. – <https://github.com/smar-git/SM-data-merging>.
Matutini, F. et al. 2021. How citizen science could improve species distribution models and their independent assessment for conservation. – Ecol. Evol. 11: 3028–3039.
Meissner, A. et al. 2012. Feeding ecology of striped dolphins, *Stenella coeruleoalba*, in the north-western Mediterranean sea based on stable isotope analyses. – J. Mar. Biol. Assoc. UK 92: 1677–1687.
Mikula, P. and Tryjanowski, P. 2016. Internet searching of bird–bird associations: a case of bee-eaters hitchhiking large African birds. – Biodivers. Observ. 7: 1–6.
Miller, D. A. W. et al. 2019. The recent past and promising future for data integration methods to estimate species' distributions. – Methods Ecol. Evol. 10: 22–37.
Monsarrat, S. et al. 2019. Accessibility maps as a tool to predict sampling bias in historical biodiversity occurrence records. – Ecography 42: 125–136.
Pace, D. S. et al. 2012. Anthropogenic food patches and association patterns of *Tursiops truncatus* at Lampedusa island, Italy. – Behav. Ecol. 23: 254–264.
Pace, D. S. et al. 2014. Foreword. – Aquat. Conserv. 24: 1–3.
Pace, D. S. et al. 2018. Habitat suitability modeling in different sperm whale social groups. – J. Wildl. Manage. 82: 1062–1073.
Pace, D. S. et al. 2019. An integrated approach for cetacean knowledge and conservation in the central Mediterranean sea using research and social media data sources. – Aquat. Conserv. 29: 1302–1323.
Pace, D. S. et al. 2021a. Capitoline dolphins: residency patterns and abundance estimate of *Tursiops truncatus* at the Tiber river estuary (Mediterranean sea). – Biology 10: 275.
Pace, D. S. et al. 2021b. Facts and outcomes of the Mediterranean short-beaked common dolphin *Delphinus delphis* workshop. – Aquat. Conserv. 31: 5–7.
Phillips, S. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.
Pulcini, M. et al. 2014. Distribution and abundance estimates of bottlenose dolphins *Tursiops truncatus* around Lampedusa island (Sicily channel, Italy): implications for their management. – J. Mar. Biol. Assoc. UK 94: 1175–1184.
Redfern, J. V. et al. 2006. Techniques for cetaceanhabitat modeling. – Mar. Ecol. Prog. Ser. 310: 271–295.
Redfern, J. V. et al. 2017. Predicting cetacean distributions in data-poor marine ecosystems. – Divers. Distrib. 23: 394–408.
Rue, H. and Held, L. 2005. Gaussian Markov random fields. Theory and applications. – Chapman & Hall/CRC.
Rue, H. et al. 2009. Approximate bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. – J. R. Stat. Soc. B Stat. Methodol. 71: 319–392.
Santamaria, C. et al. 2017. Mass processing of Sentinel-1 images for maritime surveillance. – Remote Sens. 9: 678.
Sicacha-Parada, J. et al. 2020. Accounting for spatial varying sampling effort due to accessibility in citizen science data: a case study of moose in Norway. – Spat. Stat. 42: 100446.
Simpson, D. et al. 2016. Going off grid: computationally efficient inference for log-Gaussian Cox processes. – Biometrika 103: 49–70.
Soranno, P. A. and Schimel, D. 2014. Macrosystems ecology: big data, big ecology. – Front. Ecol. Environ. 12: 3.
Stephenson, F. et al. 2020. Modelling the spatial distribution of cetaceans in New Zealand waters. – Divers. Distrib. 26: 495–516.
Thompson, S. K. and Ramsey, F. L. 1987. Detectability functions in observing spatial point processes. – Biometrics 43: 355–362.
Triossi, F. et al. 2013. Occurrence of bottlenose dolphins *Tursiops truncatus* in natural gas fields of the northwestern Adriatic sea. – Mar. Ecol. 34: 373–379.
Vella, A. et al. 2021. The conservation of the endangered mediterranean common dolphin *Delphinus delphis*: current knowledge and research priorities. – Aquat. Conserv. 31: 110–136.
Ventura, D. et al. 2015. Temporal partitioning of microhabitat use among four juvenile fish species of the genus *Diplodus* (pisces: Perciformes, Sparidae). – Mar. Ecol. 36: 1013–1032.
Yuan, Y. et al. 2017. Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. – Ann. Appl. Stat. 11: 2270–2297.