Black Scabbardfish Species Distribution: Geostatistical Inference Under Preferential Sampling

Paula Simões[1,2(✉)], M. Lucília Carvalho[3], Ivone Figueiredo[3,4], Andreia Monteiro[1], and Isabel Natário[1,5]

[1] NOVA MATH - Center for Mathematics and Applications, NOVA University of Lisbon, Lisbon, Portugal
pc.simoes@campus.fct.unl.pt
[2] Military Academy Research Center - Military University Institute (CINAMIL), Lisbon, Portugal
[3] Centre of Statistics and its Applications (CEAUL), Faculty of Sciences of the University of Lisbon, Lisbon, Portugal
[4] Portuguese Institute for Sea and Atmosphere (IPMA), Lisbon, Portugal
[5] Department of Mathematics, Faculty of Science and Technology, NOVA University of Lisbon, Caparica, Portugal

Abstract. Black Scabbardfish (BSF) is a highly prized deep-sea species that occurs in continental waters at depths greater than 800 m. It has been recognized that improving knowledge of its biodiversity and abundance along the Portuguese coast of BSF species is a scientifically and socially relevant issue, mainly due to the fact of absence of dedicated deep-water research surveys in this area, the spatial distribution of its abundance is mainly inferred from commercial deep-water longline fishery that operates along the continental slope. Black Scabbardfish (BSF) captures are modelled using a geostatistical analysis combined with a preferential sampling technique which enables to better capture the variability of the BSF captures providing a more realistic pattern of BSF distribution. This approach allows a better knowledge os BSF spatial distribution assuming that the selection of the sampling locations depends on the values of the observed variable of interest. BSF captures are jointly modeled with their locations, using a Bayesian approach and INLA methodology, considering stochastic partial differential equations (SPDE) in the geostatistical model and in the Log-Cox point process model for the locations. Several different covariates and random effects were considered. The best two fits are presented, the first including covariate depth in the intensity of the point process besides the shared spatial effect with the response, and the second fit having covariate vessel tonnage in the response adding to the shared spatial effect and covariate depth again included in the point process intensity.

Supported by national funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects PREFERENTIAL, PTDC/MAT-STA/28243/2017, UIDP/00297/2020 (Center for Mathematics and Applications) and UIDB/00006/2020(CEAUL)x.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023 Workshops, LNCS 14105, pp. 303–314, 2023.
https://doi.org/10.1007/978-3-031-37108-0_19

Keywords: Preferential sampling · Geostatistics · Point process · SPDE · INLA

1 Introduction

Black Scabbardfish (BSF) is a deep-water species that occurs in continental waters at depths greater than 800 m. On the portuguese coast BSF constitutes an important commercial resource. In the absence of dedicated deep-water research surveys in this area, the spatial distribution of its abundance is mainly inferred from commercial deep-water longline fishery operating along the continental slope.

The Portuguese Institute of Sea and Atmosphere (IPMA) provided georeferenced data on the location of the fishing hauls and the corresponding captures for a number of differently sized vessels belonging to the Black Scabbardfish (BSF) fishing fleet. The commercial longline fishing that operates along the continental slope is the main contributor to the information obtained on the spatial distribution of the abundance of this species, as dedicated deep-water research surveys in this area are scarce. This information, understood as preferentially sampled data, is used here combined with environmental covariates, to predict where the species is likely to exist, also in unsampled locations, for management and conservation purposes, in order to ensure the sustainability of commercial fisheries and protect the biodiversity of species that are of high interest for consumption.

A classical geostatistical approach was considered in a preliminary study of this species [1], using different regression models with covariates, structured and unstructured random effects, under a Bayesian approach. However, the information on the capture sites of the species should have been accounted for in the modelling process due to the preferential nature of the data, as classical geostatistical techniques may have produced biased results [5,10,15].

The objective of this research work is two-folded. In one hand, to perform species distribution modelling of the BSF data by using a geostatistical model-based method that takes preferentiability into account. Under a Bayesian approach and resorting to INLA methodology [11], using the stochastic partial differential equations (SPDE) approach to model the spatial effects that are shared by a geostatistical model for the captures and a Log-Cox point process model (LGCP) for the locations [7], BSF captures can be analysed considering several different covariates and random effects [5,10]. On the other hand, considering the present modelling of the BSF captures, predictions can be made at several potential sampling locations (unobserved locations) in order to be able to construct a survey design to improve the BSF abundance estimates in Portuguese waters. The present study contributes to the first part of the analysis, to perform species distribution modelling. This analysis is used in the second part of the work, to construct a survey design to improve the BSF abundance estimates, as in [3,13].

Considering the referred Bayesian framework, using INLA and SPDE for estimating jointly the geostatistical model for captures and the Log-Cox point

process model for the locations, BSF captures are estimated using different models. From these, both in terms of a classical geostatistical approach and further taking preferentiability into account, the two best models are presented. Within the latter models, first it is considered the inclusion of covariate depth in the intensity of the point process and of a spatial effect in the response, and for the second one the covariate group of tonnage is also incorporated in the response.

2 Species Distribution Modelling for the Black Scabbardfish

2.1 Black Scabbardfish Data

A set of 15 vessels belonging to the BSF fishing fleet, with different sizes, contributes through the Portuguese Institute of Sea and Atmosphere (IPMA) in providing georeferenced data corresponding to the BSF catches (in Kg) and respective fishing haul locations. The database also includes information in terms of tonnage and vessel identification, speed and depth at which the catch was made. The time period covers the years 2002 to 2013, however a subset of the original data was taken for this data analysis: it comprehends the years between 2009 to 2013, catches that took place from September to February on the southern fishing zone that corresponds to a latitude below than 39.3. This subset of the original data includes 733 observations. Figure 1 shows the spatial distribution of the capture points of the considered data subset. In addition to the available variables, it is also intended to consider environmental factors as covariates to perform species distribution modelling. For the present analysis the covariates depth revealed to be significant.

The original data displays skewness, with a Gamma distribution possibly being a good fit for the observed captures. In order to have an approximately Normal distribution of the response, a Box-Cox transformation of BSF captures data (Y) was carried out. The new transformed response (Y^*) as been defined according with the expression $Y^* = \frac{Y^\lambda - 1}{\lambda}$, with $\lambda = \frac{1}{2}$.

The vessel tonnage is classified as low tonnage (less than 26 tonnes) and medium Tonnage (higher than 26). 484 of observations are from the vessels are registered as low tonnage, and the remaining 249 observations are registered as medium tonnage. Depth variable has a minimum value of -1997.4 m, median value of -1298.3 m and a interquartile range of 214.0 m with a maximum value of -770.8 m.

2.2 Geostatistical Classical Model Versus Preferential Sampling Model

Let Y_i denote the georeferenced species captures for site i, $i = 1, \ldots, n$, assuming the following distribution for the observations, $Y_i \sim N(\mu_i, \sigma_o^2)$ where n is the

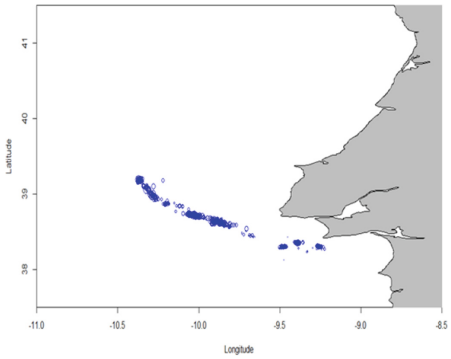

Fig. 1. Black Scabbardfish data locations.

total number of fishing hauls, σ_o^2 is the nugget effect, μ_i is the response mean. The classical geostatistical model is given by:

$$\begin{gathered} Y_i|(S, \boldsymbol{X} = \boldsymbol{x})] = \mu_i + e_i = \beta_0 + \textstyle\sum_{m=1}^{M} \beta_m C_{mi} + S(x_i) + e_i, \\ \boldsymbol{e} \sim N(\boldsymbol{0}, \sigma_o^2), \end{gathered} \tag{1}$$

where $\boldsymbol{C} = (C_1, \ldots, C_M)$ corresponds to the considered covariates in the study, $\boldsymbol{\beta}$ is the corresponding fixed effects, essentially environmental factors and available information about the vessels. The spatial effect of the model is given by S, modelled as Latent Gaussian Field (GF), a stationary Gaussian Process with mean zero, variance σ^2, and a Matérn correlation function, with correlation shape parameter fixed $k = 1.5$, correlation range ϕ and nugget variance τ^2, [2,7].

The spatial effects are approximated by the SPDE approach, $S(x) = \sum_{g=1}^{G} A_{ig}\widetilde{S}_g$. $\widetilde{S}_g$ are zero mean Gaussian distributed weights; A_{ig} is the generic element of the sparse $n \times G$ matrix A that maps the Gaussian Markov Random Field (GMRF), projecting the $\widetilde{S}$ defined on the G vertices of a triangulation of the area under study (forming a mesh) to the n locations $(x_1, ..., x_n)$ [1,2].

For β_m Normal non informative priors are assumed, $\beta_m \sim N(0, 0.001^{-1})$. In terms of the marginal standard deviation σ and the range r of the Matérn field, the penalise complexity priors (PC prior) are considered here. Based on the work of Fuglstad (2019) [6], that extends the work of Simpson (2017) [12], a joint prior for the range and for marginal variance of one-dimensional, two-dimensional or three-dimensional Matérn Gaussian Random Field with fixed smoothness was developed. Essentially it corresponds to the construction of a prior for the Matérn covariance function of a considered GMRF that is independent of the observation process. The prior distribution is weakly informative and penalizes complexity by reducing the range to infinity and the marginal variance towards zero [6]. This approach stands out by the fact that is a way to overcome the problem of overfitting, featuring that the prior distribution do not depend

on the spatial design, improving predictive performance [6]. In this work, the hyperparameters are select in order to satisfy:

PC prior for the range: $P[r < r_0] = \alpha_1 \rightarrow P[r < 30] = 0.2$;

PC prior for the standard deviation: $P[\sigma > \sigma_0] = \alpha_2 \rightarrow P[\sigma > 10] = 0.01$.

With these prior distributions a standard deviation greater than 10 is considered as well as a range less than 30.

The presented spatial model (1) corresponds to a classical geostatistical model considered under a Bayesian approach. In order to take into consideration the preferential nature of these data, it is also needed to consider the fishing sites of the data in the modelling process. That is a preferential sampling model should be considered, wherein captures of a species of interest are jointly modeled with their locations. The observed locations $(x_1, ..., x_n)$, are assumed to form a non-homogeneous Poisson process. Log-Gaussian Cox Process (LGCP) are a specific class of Cox processes in which the logarithm of the intensity surface is a Gaussian Random field. Through the underlying spatial field, S, defined as before, the observed captures $Y = (Y_1, ..., Y_n)$ are related to the intensity of the point process. Based on a trend function that may depend on covariates, the species distribution model is described. This preferentially sampling model share information in what concerns to the mean captures and sampling intensity (the number of points per unit area), on the observed species captures and on the point process [7,9,10].

$$\begin{aligned} &Y_i|(S, \boldsymbol{X} = \boldsymbol{x})] = \beta_0^y + \beta^y S(x_i) + e_i \\ &\boldsymbol{e} \sim N(0, \sigma_o^2) \\ &\boldsymbol{X}|S \sim P(\lambda) \\ &\lambda_i|S = \exp(\beta_0^{pp} + S(x_i)) \end{aligned} \tag{2}$$

where β_0^y is the intercept for the model of the captures, β_0^{pp} is the intercept for the intensity of the LGCP, β^y is a weight on the shared spatial random field effect. If more observation sites correspond to higher captures, a value of $\beta^y > 0$ is expected; when $\beta^y = 0$ the model has no preferential sampling; when there are more observation sites for lower captures a value of $\beta^y < 0$ is estimated.

Inference in made considering the SPDE approach for geostatistical data and Integrated Nested Laplace Approximation (INLA) methodology. An approximate stochastic weak solution of a Stochastic Partial Differential Equation is used, that is a continuous Gaussian field with a Matérn covariance structure, [8,11]. The triangulation of the considered spatial domains, through the constrained refined Delaunay triagulation, is required in the first stage. A mesh that covers the study region is built, see Fig. 2. It was also considered penalized complexity priors (PC Priors) for the model parameters, range r and marginal standard deviation σ, of the spatial effect, ($P[r < 30] = 0.2$, $P[\sigma > 10] = 0.01$) for the SPDE approach [6,13].

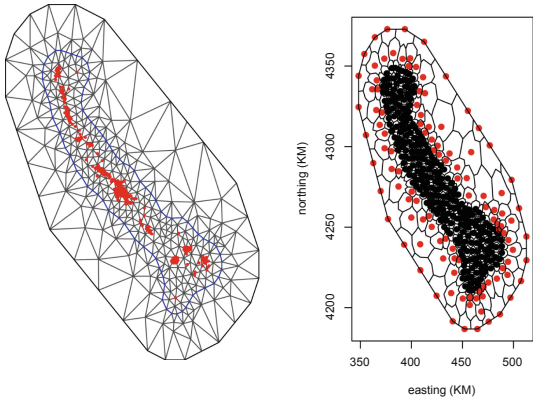

Fig. 2. Mesh constructed for BSF Data (left) and Voronoy polygons used to make inference for LGCP (right).

3 Geostatistical Inference Under Preferential Sampling

The BSF captures are analysed, considering different models, both for the classic geostatistical approach and considering the joint model of the geostatistical data and the points location [2,7]. The two best fits, both for the classical geostatistical point of view and considering the preferential sampling approach, are presented. Several different covariates and random effects were considered. The first model presented includes covariate depth in the intensity of the point process along with the shared spatial effect with the response, and the second model includes covariate vessel tonnage in the response along with the shared spatial effect and covariate depth again included in the point process intensity.

Inference is made mainly through the package R-INLA [8,11].

3.1 Classical Models

Model 1C: Including Covariate *D* Depth (Fixed Effect) and a Spatial Random Effect for Modelling Captures

$$\begin{aligned} &Y_i|(S, \boldsymbol{X} = \boldsymbol{x})] = \beta_0 + \beta_1 D_i + S(x_i) + e_i, \\ &\boldsymbol{e} \sim N(0, \sigma_o^2), \end{aligned} \tag{3}$$

where β_0 is the intercept, β_1 is the depth fixed effect and S is the random spatial effect.

Figure 3 shows the posteriori predicted mean of the spatial random effect and the corresponding standard deviation for model 1C, as well as the Box-Cox transformation of the BSF captures data observations.

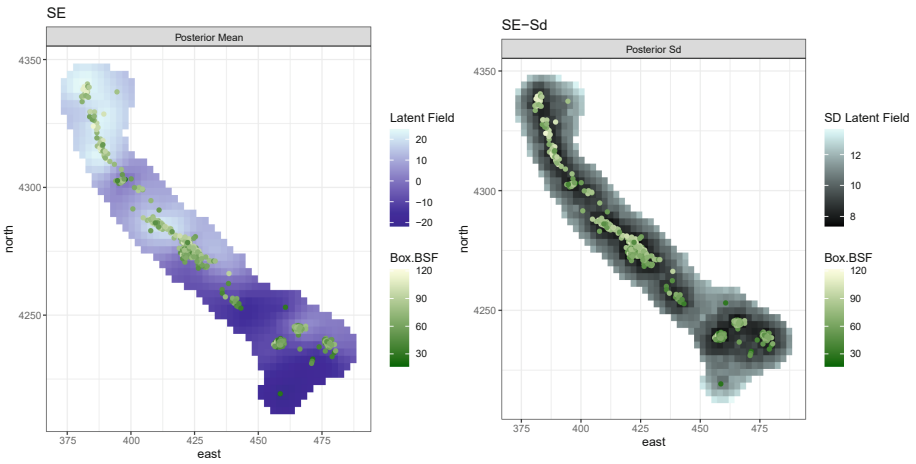

Fig. 3. Posterior predicted mean and standard deviation of the spatial effect, model 1C, and the Box-Cox transformation of the BSF captures data observations.

Model 2C: Including Covariate Group of Tonnage, *PRT* (Fixed Effects) and a Spatial Random Effect for Modelling Captures

$$\begin{aligned} &Y_i|(S, \boldsymbol{X} = \boldsymbol{x})] = \beta_0 + \textstyle\sum_{m=2}^{M} \beta_m PRT_{mi} + S(x_i) + e_i, \\ &\boldsymbol{e} \sim N(0, \sigma_o^2), \end{aligned} \tag{4}$$

where β_0 is the intercept, β_m are the group tonnage fixed effects and S is the spatial random effect.

For model 2C the posteriori predicted mean of the spatial random effect and the corresponding standard deviation, as well as the Box-Cox transformation of BSF captures data observations can de seen in Fig. 4.

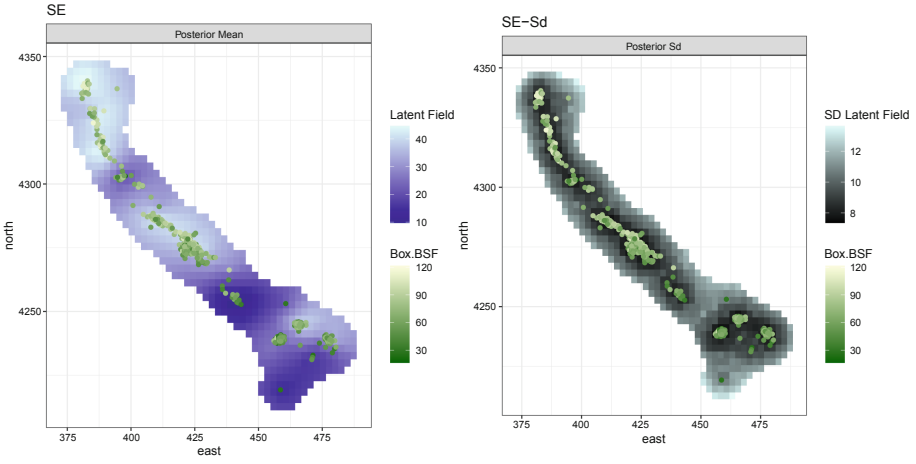

Fig. 4. Posterior predicted mean and standard deviation of the spatial effect, model 2C, and the Box-Cox transformation of BSF captures data observations.

3.2 Preferential Models

Model 1P: Point Process with Covariate Depth (D) on Intensity and Sharing a Random Spatial Effect with Captures

$$\begin{aligned} &Y_i|(S, \boldsymbol{X} = \boldsymbol{x})] = \beta_0^y + \beta^y S(x_i) + e_i, \\ &\boldsymbol{e} \sim N(0, \sigma_o^2), \\ &\lambda_i|S = \exp(\beta_0^{pp} + \beta_1^{pp} D_i + S(x_i)), \end{aligned} \tag{5}$$

where β_0 is the captures model intercept, β_0^{pp} is the point process intercept, β_1 is the depth fixed effect and S is the shared spatial random effect. The observed locations are modelled by a Log-Gaussian Cox Process (LGCP), $X|S$.

Figure 5 shows the posteriori predicted mean of the spatial random effect and the corresponding standard deviation, as well as the Box-Cox transformation of BSF captures data observations for model 1P. This model has an estimated value of β^y of **1.237**, greater than 0, meaning that the BSF captures, are **higher** where there are **more** observation locations.

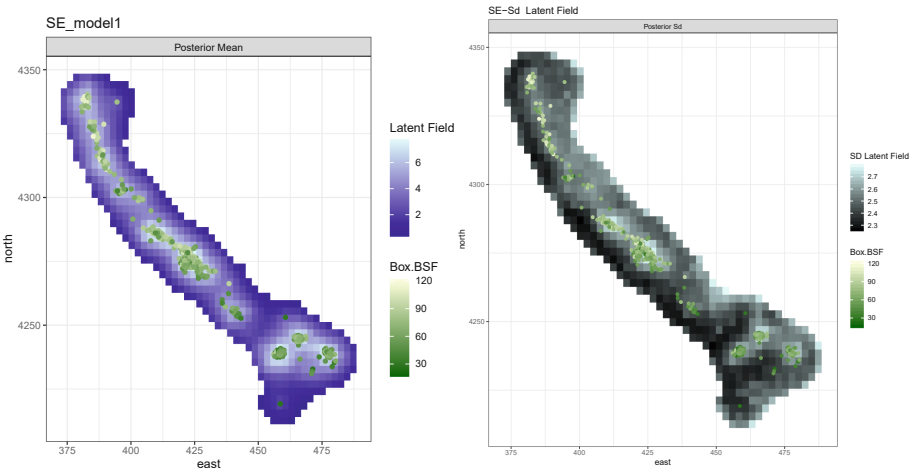

Fig. 5. Posterior predicted mean and standard deviation of the spatial effect, model 1P, and the Box-Cox transformation of BSF captures data observations.

Model 2P: Including Covariate *PRT*, Group of Tonnage (Fixed Effect), Point Process with Covariate Depth (D) on Intensity and Sharing a Random Spatial Effect with the Captures

$$\begin{aligned} &Y_i|(S, \boldsymbol{X} = \boldsymbol{x})] = \beta_0^y + \beta_1^y PRT_i + \beta^y S(x_i) + e_i, \\ &\boldsymbol{e} \sim N(0, \sigma_o^2), \\ &\lambda_i|S = \exp(\beta_0^{pp} + \beta_1^{pp} D_i + S(x_i)) \end{aligned} \tag{6}$$

where β_0 is the intercept for the model of the captures, β_0^{pp} is the point process intercept, β_1 is the depth fixed effect and S is the shared spatial random effect. The observed locations are modelled by a LGCP, $X|S$.

For model 2P the posteriori predicted mean of the spatial random effect and the corresponding standard deviation, as well as the Box-Cox transformation of the BSF captures data observations can de seen in Fig. 6. This model has an estimated value of β^y of **0.228**, greater than 0, meaning that the response values are **higher** where there are **more** observation locations.

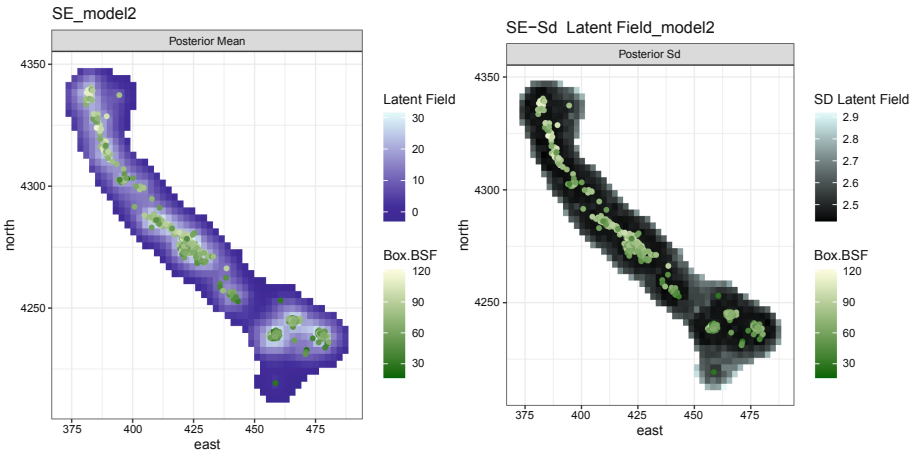

Fig. 6. Posterior predicted mean and standard deviation of the spatial random effect, model 2P, and the Box-Cox transformation of BSF captures data observations.

3.3 Conclusion

Parameter estimates for two selected models, both in terms of the classical and preferential approaches, are summarized in Tables 1 and 2, respectively. Using the Deviance Information Criterion (DIC) and the Watanabe-Akaike Information Criterion (WAIC), well-known Bayesian model choice criterion for comparing models, these models were selected, from several other fits, through their predictive accuracy [14].

Table 1. Parameter estimates for the classical approach, model 1C and model 2C.

Model	β_0^y	D	PRT medium	r (range)	σ (StDev)	τ_o (Prec Gaussian)	DIC	WAIC
Model 1C	62.432	−0.001	-	37.791	14.448	0.006	5893.07	5898.09
Model 2C	29.654	-	14.761	38.381	12.579	0.006	5866.95	5871.39

Table 2. Parameter estimates for the preferential approach, model 1C and model 2C.

Model	β_0^{pp}	D	β_0^y	PRT medium	r (range)	σ (StDev)	τ_o (Prec Gaussian)	β^y	DIC	WAIC
Model 1P	−10.710	−0.002	62.253	-	39,128	3.784	0.003	1.237	7459,92	8850.97
Model 1P	−6.680	0.001	32.914	24.304	25,573	11.999	0.005	0.228	7269,67	8650.00

The values for DIC and WAIC in each approach (classical and preferential) are very similar. In spite of this, for the classical approach, the model including covariates PRT and group of tonnage (fixed effect) and a spatial random effect for modelling captures has the lowest values. For the preferential approach, the model including covariates PRT and group of tonnage (fixed effect) for the captures, and covariate depth (D) for the point process intensity, sharing with captures the random spatial effect seems to be the best model to fit the data.

4 Discussion

BSF captures were modelled considering several different covariates and spatial random effects, considering a joint model of the geostatistical data and the points location and also via a classical geostatistical model, taken under a Bayesian approach. The preferential sampling model better captures the variability of the BSF captures providing a more realistic pattern of the BSF distribution. In the presence of preferentially of the sampled data (the sampling locations depends on the values of the observed variable of interest) this approach allows a better understanding of the spatial distribution of BSF. This information combined with environmental covariates, can now be used to predict where the species is likely to exist, also in unsampled locations, for management and conservation purposes, in order to ensure the sustainability of commercial fisheries and protect the biodiversity of this species that is of high interest for consumption. The study could be more broadly extended in such a way as to include important ecological factors, which may be important in the estimation, or by allowing a term for the time effects, in terms of a spatio-temporal approach [2,9].

Taking into account the present modelling of the BSF captures, predictions can be made at several potential sampling locations (unobserved locations) in order to construct a survey design to improve the BSF abundance estimates, which represents the second part of the work [13]. These locations can be obtained through the triangulation of the considered spatial domain, taking into account the chosen mesh. The estimated values for BSF captures in their spatial prediction coordinates can be obtained considering both approaches (classical vs preferential), what might constitute an important issue, for future approach, in order to be able to analyse the effect of preferentiality on the choice of new sampling locations, for sampling design purposes, and its influence in the sampling design choice.

In terms of future work, the aim is thus to implement geostatistical sampling design strategies for preferentially sampled data, addressing the problem

of designing for spatial prediction purposes when the covariance function is not known and has to be estimated from the same data. Within this approach, different design classes could be investigated, namely random, simple inhibitory and adaptative geostatistical samplig designs [3,4], in order to reach an optimal survey BSF design. Another important issue to be further considered is the development of a new sampling design without transforming the captures BSF, that also enables a realistic pattern of BSF captures distribution estimates. This will require that the Geostatistical preferential model to be considered for the BSF captures assumes a Gamma distribution instead, as well as the follow up with suitable design strategies for BSF captures under preferential sampling with this Gamma distribution.

These latter issues are considered pertinent to the ongoing mission of improving knowledge about the abundance of certain fish species (BSF or others), which is a extremely determining factor to support, from a national as well as international perspective, the sustainability of commercial fisheries and safeguard the biodiversity of species of significant consumption interest.

Acknowledgments. This work is funded by national funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects PREFERENTIAL, PTDC/MAT-STA/28243/2017, UIDP/00297/2020 (Center for Mathematics and Applications) and UIDB/00006/2020(CEAUL).

References

1. André, L.M., Figueiredo, I., Carvalho, M.L., Simões, P., Natário, I.: Spatial modelling of black scabbardfish fishery off the Portuguese coast. In: Gervasi, O., et al. (eds.) ICCSA 2020. LNCS, vol. 12249, pp. 332–344. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58799-4_25
2. Blangiardo, M., Cameletti, M., Baio, G., Rue, H.: Spatial and spatio-temporal models with R-INLA. Spat. Spatio-Temporal Epidemiol. **4**, 33–49 (2013)
3. Chipeta, M., Terlouw, D., Phiri, K., Diggle, P.: Adaptive geostatistical design and analysis for prevalence surveys. Spat. Stat. **15**, 70–84 (2016)
4. Chipeta, M., et al.: Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure. Environmetrics **28**(1), e2425 (2017)
5. Diggle, P., Menezes, R., Su, T.: Geostatistical inference under preferential sampling. J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) **59**(2), 191–232 (2010)
6. Fuglstad, G.A., Simpson, D., Lindgren, F., Rue, H.: Constructing priors that penalize the complexity of Gaussian random fields. J. Am. Stat. Assoc. **114**(525), 445–452 (2019)
7. Krainski, E.T., et al.: Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA. CRC Press, Boca Raton (2018)
8. Lindgren, F., Lindström, J., Rue, H.: An explicit link between Gaussian fields and Gaussian Markov random fields: the SPDE approach. Centre for Mathematical Sciences, Faculty of Engineering, Lund University, Mathematical Statistics (2010)
9. Martínez-Minaya, J., Cameletti, M., Conesa, D., Pennino, M.G.: Species distribution modeling: a statistical review with focus in spatio-temporal issues. Stoch. Env. Res. Risk Assess. **32**(11), 3227–3244 (2018). https://doi.org/10.1007/s00477-018-1548-7

10. Pennino, M., et al.: Accounting for preferential sampling in species distribution models. Ecol. Evol. **9**(1), 653–663 (2019)
11. Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc.: Ser. b (Stat. Methodol.) **71**(2), 319–392 (2009)
12. Simpson, D., Rue, H., Riebler, A., Martins, T.G., Sorbye, S.H.: Penalising model component complexity: a principled, practical approach to constructing priors. Stat. Sci. **32**(1), 1–28 (2017)
13. Simões, P., Carvalho, M.L., Figueiredo, I., Monteiro, A., Natário, I.: Geostatistical sampling designs under preferential sampling for black scabbardfish. In: Bispo, R., Henriques-Rodrigues, L., Alpizar-Jara, R., de Carvalho, M. (eds.) SPE 2021. Springer Proceedings in Mathematics and Statistics, vol. 398, pp. 137–151. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-12766-3_11
14. Spiegelhalter, D., Best, G., Carlin, P., Van Der Linde, A.: Bayesian measures of model complexity and fit. J. R. Stat. Soc.: Ser. b (Stat. Methodol.) **64**(4), 583–639 (2002)
15. Watson, J., Zidek, J.V., Shaddick, G.: A general theory for preferential sampling in environmental networks. Ann. Appl. Stat. **13**(4), 2662–2700 (2019)