SC/68A/EM/03

Spatial density modelling using statistical and machine learning methods and its application to the Antarctic krill

Fang Lu, Christian Reiss, George Watters, Hiroto Murase and Toshihide Kitakado

INTERNATIONAL
WHALING COMMISSION

Spatial density modelling using statistical and machine learning methods and its application to the Antarctic krill

Fang Lu[1#], Christian Reiss[2], George Watters[2], Hiroto Murase[1] and Toshihide Kitakado[1]*

[1] Tokyo University of Marine Science and Technology
[2] South West Fishery Center, NOAA
*Corresponding author's address: kitakado@kaiyodai.ac.jp, 5-7, Konan 4, Minato, Tokyo 108-8477, Japan
[#]Currently Qingdao Tao Ran Environmental Science and Technology Co. Ltd

ABSTRACT

As a key species linking primary producers to the higher tropic levels in the Antarctic ecosystem, the Antarctic krill (*Euphasia superba*) plays an important role in the Antarctic ecosystem. Therefore, knowing the plausible spatial distribution of the krill will be useful for the resource management and conservation in this area. Species distribution models (SDMs) can help predict the spatial species density by quantifying the relationship between the observed species distribution and its influencing factors. In general, although both statistical models and machine learning methods can be applied as SDMs, there is a still open question of how the estimation performance of those SDMs for the Antarctic krill is. To address this question, we conducted simulation studies for six different SDMs under two different survey-designs with zig-zag-shaped and tooth-shaped track lines, in order to assess the estimation performance of each model for krill distribution and mean density. As the procedure, we first conditioned two different density distribution of krill in this region by using actual krill density observation data. Using the assumed true spatial density surfaces, we repeatedly generated simulation data under the two designs, and then applied six SDMs, two statistical models and four machine learning methods, to the data. As performance measures, the mean squared error of predicted surface (MSE), relative bias and root mean squared error (RMSE) for the mean density were used. As a result, machine leaning methods were proven to have higher and more reliable prediction abilities than traditional statistical models, especially random forests (RF) and boosted regression trees (BRT) were revealed to be the most reliable methods in this study. In addition, the zigzag-shaped and tooth-shaped designs are found to have comparable performances, and both of them can be applied in the krill field survey.

Keywords: *Antarctic krill, line transect, machine learning, simulation, species distribution model, survey design*

INTRODUCTION

The Antarctic krill (*Euphasia superba*) plays a significant role in the management of the fishery resources in the Antarctic Ocean since it links the primary producers with the higher trophic species such as marine mammals. As in the cases of other fishery populations, understanding of the abundance level and spatial distribution contributes to the management of this species. For this purpose, some field surveys in the Antarctic Ocean are conducted for the data collection. However, the costs of these surveys are pretty high, and therefore the development of better survey designs are of great interest.

The line transect methods with acoustic observation has been a primary platform for the krill abundance survey. Usually a tooth-shaped design of track lines has been used for krill survey while a zigzag-shaped design is used in most of cetacean surveys. To arrange the krill survey in conjunction with the cetacean survey, it is questioned if the zigzag shaped survey design is also feasible and can produce comparable estimates of spatial density distribution as well as mean density over the survey area with those by the tooth shaped design.

As for the data analyses for estimating spatial density distribution of krill, species distribution models (SDMs) can be used to quantify the relationship between the observed species distribution and the influencing factors such as sea surface temperature. The model can also predict the potential distribution of this species based on this relationship. As traditional statistical techniques, "generalized linear models" (GLMs; McCullagh and Nelder 1989) and "generalized additive models" (GAMs; Hastie and Tibshirani 1990) have been used as SDMs and well recognized in the field of ecological modeling for their ability to handle a variety of error structures. Recently, a wide variety of machine learning methods/techniques have also been applied to species distribution modeling due to their ability to handle large quantities of data, to understand the complex relationship between response and predictors, and to generate relatively reliable predictions (Moisen *et al.* 2006).

In this paper, we conducted simulation studies to provide a comprehensive comparison of the effectiveness of the two survey designs used for the krill resource estimation i.e. tooth and zigzag. We also compared the estimation performances between the two traditional statistical models and four machine learning methods which were used to predict the spatial density distributions of krill.

MATERIALS AND METHODS

Here, the region of interest, as well as the response and predictor variables used for developing the SDM are described. The six models used in this analysis are described briefly, including GLM, GAM, least absolute shrinkage and selection operator (Lasso; Tibshirani (1996), support vector machine (SVM; Basak *et al*. 2007), random forests (RF; Breiman 2001), boosted regression trees (BRT; Friedman 2001).

Data description

This section describes the data used in this study. This includes a description of the study region, the response variable and the predictor variables used in this analysis.

Study region

The study area is located between 53°-65°W and 59°-65°S, to the north of the Antarctic Peninsula, where the annual field studies of the U.S. Antarctic Marie Living Resources (U.S. AMLR) are conducted (Fig. 1).

Response variable ($\boldsymbol{y}$)

To develop the SDM for predicting the krill density distribution, we take the density value of krill as the response variable, indicated by den in this study, which is the objective we try to model by understanding the link with its influencing factors, such as the temperature, etc. The krill density is estimated from the acoustic surveys. These surveys were conducted by the U.S. AMLR Program in 2011 (mid-January to early March), and averaged over 1-nmi intervals (Fig. 2). In total, 1439 density values were recorded, with the maximum of 1879.45 g/m^2 and mean value of 60.91 g/m^2. When the log of the observed density values (which come from survey data) are taken, the resulting distribution is approximately normal in shape (Fig. 3).

Predictor variables

Predictor variables are the factors that may have influence on the variation of the response variable. The single models (GLM, GAM, Lasso with GLM, SVM, RF and BRT) are run with eight predictor variables: *lon* (longitude), *lat* (latitude), *SST* (sea surface temperature), *sal* (salinity), *chla* (the concentration of chlorophyll a), *zoo* (the concentration of zooplankton), *depth* (depth), and *distance* (distance from land). All the eight predictors are continuous variables. The information for the coordinates (lon and lat), SST, sal, chla and zoo data were collected during the AMLR survey, while the depth data was downloaded from *https://www.gebco.net/.* The distance from land data was calculated using the "*distm*" function in the package '*geosphere*' in R, with the land edge coordinates used. The distance from land values indicate the shortest distance between the sampling stations used in survey and lands.

Models

The general relationship between the response variable and the predictor variables in this study can be described as follows:

$$\log(density) \sim lon + lat + SST + sal + chla + zoo + depth + distance \tag{1}$$

where the left side is the modelling objectives, the response variable, while the predictor variables on the right side are used to describe the response variable.

The relationship between the spatial density distribution of krill (the response variable) and the predictor variables is described using six models: GLM, GAM, Lasso, SVM, RF, BRT. The main features and the details of the settings for running in R are summarized in Table 1.

On the basis of linear regression model, GLM is the extension of the ordinary linear model that allows for response variables have various types of error distributions other than a normal distribution. As a traditional statistical method, the GLM has been recognized within the scope of ecological modelling e.g. predicting the plant distribution in the Spring Mountains of Nevada, USA (Guisan *et al.* 1999). GAM is the semi-parametric extension of GLM (Yee and Mitchell 1991), maintaining the additivity of the modelling approach, and allowing for non-linear functions in the predictor variables. As a result, GAM has been frequently used in the prediction of species distribution because of its ability to handle non-linear relationships. Some examples of its application include the explanation of the species distribution with respect to climate (Yee and Mitchell 1991), and predicting the occurrence of fern species in New Zealand (Zaniewski *et al.* 2002).

Machine learning methods have also been applied in the spatial modelling scope recently, with their ability to understand the complex relationship between response and predictors been recognized. Lasso has the ability to yield sparse models by selecting the meaningful predictors, which can avoid the overfitting and lead to more interpretable models. SVM can provide a more tolerable error bound by minimizing the convex objective function, a combination of a loss function with a regularization term, to generate more general performance (Basak *et al.* 2007). RF is an enhancement technique that collects many deep regression trees using bagging, and averages the results of all trees to get a more accurate and stable prediction. BRT is also an enhancement of regression trees, with the trees are grown to the residuals in an adaptive manner as opposed to being identically distributed as is in RF. The additive procedure of BRT merges all the shallow trees to make a stronger learner.

Conditioning of models for simulation

This section details how the simulation of this study was conducted. It consists of two parts: 1) the conditioning of the true krill density distribution surface 2) the simulation work of the krill spatial modelling.

Conditioning of krill density surface

For the simulation of the krill survey designs, the krill density surface of the overall area is conditioned by using actual data. There are six candidate surfaces in total and the best performance models on the conditioning will be selected for the conditioning.
The evaluation measurement for the conditioning ability of each model is the mean value of the residual sum of squares (RSS),

$$mean\ RSS = \frac{1}{n}\sum_{i=1}^{n}(\log y_i - \log \mu(\mathrm{x}_i))^2 \qquad (2)$$

where n is the row number of the data set used in the corresponding procedure, y_i indicates the ith density observation, $\mu(\mathrm{x}_i)$ represents the ith prediction from the single model. All the calculations are performed in log-scale.

The conditioning performance of each model is evaluated using three different evaluation procedures: fitting the models using the total data set, 10-fold cross-validation, spatial cross-validation.

Fitting using total data set

For this procedure, the total data set is used to calibrate the models, and the models are evaluated by the same data without density value. This procedure provides the most information to the evaluated models compared to the previous evaluation procedures and is expected to produce the lowest mean RSS.

10-fold cross-validation

K-fold cross validation is a resampling approach used to assess the robustness of a model (Van Houwelingen and Le Cessie 1990). In this study, the 10-fold cross validation is applied, where the total data set is randomly divided into ten independent partitions. Nine of these partitions are used to calibrate the models, and the remaining one partition is used to evaluate the model's performance. The above procedure is repeated ten times, till each partition has been used once as an evaluation set.

Spatial cross-validation

Similar to the 10-fold cross-validation, the data set is divided into ten partitions by dividing the study area into 10*10 spatial grids. The data covered in the randomly sampled (without replacement) ninety spatial grids are used as calibration data, which occupies ninety percent of the total data set, while the data in the remaining ten grids are used to evaluate the model's performance. The above procedure is also repeated ten times, till each partition has been used once as an evaluation set.

The selection of the "best" models for conditioning is based on the results of RSS in the above procedures. The density surfaces conditioned by the best performed models will be considered as the real krill density distributions to be used later in the simulation work.

Generation of simulation data set

Following the conditioning of the real density distribution surface, the density observations in the simulation data sets are generated from the conditioned density surface using the below distribution, with CV values equal to 10%, 30% and 50% respectively:

$$\log(density) \sim N\left(\log\mu - \frac{1}{2}\log(1+\mathrm{CV}^2), \log(1+\mathrm{CV}^2)\right) \quad (3)$$

One hundred iterations are generated from each distribution with only the CV being different. Thus, in total, three hundred iterations are generated from each conditioned density surface for the simulation work outlined below.

Simulation

Simulation of two survey designs

To compare the effectiveness of the zigzag and tooth survey designs for krill resource estimation, the simulation work on the study area is conducted by mimicking the real survey situation. Fig. 4 is a graphical representation of the survey routes in one iteration. The blue dashed lines indicate the effective survey routes of tooth design, while the red dashed lines indicate the effective zigzag survey routes. The intervals among the sample stations and the total effective survey efforts of the two designs are set to be roughly equal, with 689 stations set in the zigzag design, and 696 for the tooth design. Both the krill density and the information on the predictors are available on the sampled stations along the survey routes. In the real situation, the locations of survey routes can be influenced by many factors, therefore the locations tend to be unfixed. To mimic the real situation, we simulate the survey by randomly set the survey location on the vertical direction for each iteration. The shape of survey routes, sample intervals and sample sizes are kept same, with only the location of survey changing.

Comparison of the predictive abilities of the models

Following the data collection from the "survey", the models are calibrated with these data in order to describe the relationship between the krill density distribution and the predictor variables. The calibrated models are then applied in order to predict the probable krill density distribution in the study area, based on the relationship that was found during the calibration process. The predictive abilities of the single models are evaluated using several measures including:

1) Mean squared error (MSE): evaluates the pointwise predictions at each prediction point,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(log\, y_i - log\, \mu(x_i))^{2} \quad (4)$$

where n is the number of prediction points in a single model, y_i indicates the ith true density value, $\mu(\mathrm{x}_i)$ represents the ith predicted density value from a single model. All the calculations are performed in log-scale.

2) Root mean squared error (RMSE): evaluates the predicted mean density of predicted density distribution surface provided by each model,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(\overline{den}_{pre-i} - \overline{den}_{true}\right)^2}{n}} \qquad (2.5)$$

where n is the number of predicted density distribution surfaces for each CV value, being equal to 100 in this study. $\overline{den}_{pre-i}$ indicates the predicted mean density value provided by the ith prediction, $\overline{den}_{true}$ is a fixed value representing the true mean density value from the conditioned density distribution surface from the best performing candidate model.

3) Relative bias to the true mean density: evaluates the predicted mean density of predicted density distribution surface provided by each model,

$$Bias = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\overline{den}_{pre-i} - \overline{den}_{true}}{\overline{den}_{true}}\right) \qquad (2.6)$$

where the meaning of the symbols in this formula is same as the above descriptions in *eq.* 2.5.

RESULTS

Results of conditioning

The evaluation results for the conditioning performance are summarized in Figs. 5-8, illustrating the residual sum of squares of each model in three evaluating procedures.

Fitting using total data set

In the test with total data set, RF provided significantly lower RSS than others (Fig. 5). This can be confirmed by checking the following scatter plots (Fig. 6), illustrating the correlations between the fitting results and the density observations are plotted in this figure, which shows that the predicted values given by RF were very close to the observed values.

10-fold cross-validation

The boxplot for the RSS of 10-fold cross-validation shows that RF can lead to the most accurate prediction on the evaluation data set, while BRT also had good prediction accuracy and stable performance (Fig. 7). The outliers in the predictions provided by GAM and SVM suggest the high variation of their performance.

Spatial cross-validation

In spatial cross-validation, the RSS has lower variation among the models (Fig. 8), but generally higher RSS values were generated compared with 10-fold cross-validation. RF still performed best among the six models according to the results, while BRT also can be seen to have good performances.

According to the results described above, it can be seen that RF and BRT can provide robust performances for conditioning than other models, suggesting their promising estimations of the krill density distribution. As a result, both the simulations on the density surfaces conditioned by RF and BRT were conducted, with each of them being considered to be the true density distribution of krill in this region (Fig. 9).

Results of simulation

To check the predictive accuracy of these models at each prediction point, we compared each prediction and its corresponding true density value at each point by calculating the pointwise MSE of each model (Figs. 10 and 11). GLM and GAM provided higher MSE with some extreme values in the predictions than machine learning methods. RF and BRT showed highest predictive accuracies for the density distribution across two conditioned surfaces, two survey designs as well as all CV values according to MSE values. SVM also led to good prediction. As for the comparison of two survey designs, the difference between the tooth and zigzag design is not significant, except for the prediction provided by GLM and GAM showing the preference for tooth design.

To see if the models can correctly estimate the abundance level in this region by predicting, we took the mean density value as an index of the stock abundance in this study. The predicted mean density was compared across models with the true mean density value conditioned by RF (Fig. 12) or BRT (Fig. 13) respectively. GLM and GAM provided high variant predictions with extreme values (out of the range of this figure), while the results demonstrate the comparable predictions on mean density provided by machine learning methods. As for the better choice of survey designs, no one of the two showed absolutely advantages than the other, especially when machine learning methods were used as the prediction models.

Some evaluation measures of the predictions on mean density value for both RF and BRT surfaces were computed and summarized in Tables 2-7. Traditional statistical methods GLM and GAM led to the predictions of mean density with high mean relative bias and RMSE, suggesting the extreme values in their predictions, which agrees well with the above described results. However, when the median value of the relative bias was taken, GLM and GAM also showed comparable performances with the other methods, which can be seen from the simulation results of RF surface (Tables 3, 4). Compared to statistical models, machine learning methods were proven to be stable through this simulation. In the simulation with RF surface, lasso provided promising performance for the estimation of the mean density value, RF, BRT and SVM were found to have competitive performances as well. However, in the simulation with BRT surface, statistical models and two of the machine learning methods, lasso and SVM were found to have difficulties in predicting the mean density of krill in this area, while RF and BRT showed superior performances and generate the predictions with much lower relative biases compared to the others. As for the survey designs, in this study, the two survey designs of the track lines showed similar performance in the krill spatial modelling, without one of them being significantly better than the other. . As for the possible combination, according to Tables 2-7, it can be found that GLM tended to be more accurate when tooth design is used, whereas GAM showed the preference for zigzag design. For the machine learning methods, Lasso and SVM usually performed better with tooth design compared to the zigzag design, on the contrary the later one worked well when BRT was used as the prediction model, while both two designs worked well with RF.

To make the comparison between prediction and true density distribution more intuitive, the predicted density distribution surfaces (one iteration randomly selected) provided by each model are depicted in Figs.14-19. Generally, RF, SVM, BRT and GAM can catch the main features of the true distribution, whereas Lasso and GLM failed to predict the density distribution of krill. The predictions of RF and BRT were smoother compared with those of SVM and GAM, with some sharp decreases among hot spots predicted in the prediction of SVM, also on the edge of the predictions provided by GAM.

As a result, we can conclude that:
1) Traditional statistical models GLM, GAM can have competitive performances if the median value is evaluated, but the variance of their predictions is high;
2) Machine learning methods have stable and reliable prediction ability across all the evaluations, especially RF and BRT are proven to be the most reliable methods in this study;
3) Zigzag design and tooth design are revealed to have comparable performance in the krill spatial modelling, and both of them can be applied in the krill field survey.

DISCUSSION

Known that krill is an important trophic connection in the Antarctic ecosystem, the findings of this study will not only be helpful to the estimation and management of the krill resource itself, but also to the other related species such as its predators, whales. Given the predation happening between krill and whales, the knowledge of probable krill density distribution can provide reliable information to the prediction of potential whale distribution, especially when the survey is not able to be conducted because of the expenditure of the surveys and geology barriers of the habitats, etc.

Zigzag design has been widely used in the whale field surveys, where the individual number of whales can be counted by the observers along the survey routes, while the survey design in tooth shape has been used in krill field survey. To investigate the practicality of incorporating the krill survey into the cetacean survey, this study made effort on the comparison of the effectiveness of these two types of survey designs in krill spatial modelling, and the results showed that neither of the design options have clear advantage across the simulation, thus both of them can be applied in the krill surveys.

This study also provided the considerable information of the model prediction ability by doing the comprehensive comparison of the statistical methods and machine leaning techniques. In the conditioning process, the superior performance of RF showed its promising conditioning ability, which may also cause some concerns about its overfitting to the data, however the performance in the two types of cross-validation can provide convincing evidence that RF can prevent the overfitting effectively, which may benefit from both the enhancement of the deep trees, and the randomness contained in the RF algorithm. In the simulation of the krill spatial modelling, machine learning methods, especially BRT and RF provided robust performances across the predictions of both density distribution and mean density, whereas the traditional statistical models including GLM, GAM show relatively unstable performances. The high variance in the predictions provided by GLM and GAM can be probably accounted for their data-based nature, which shows high sensitivity to the simulation data through the iterations. In addition, possibly the linear structures of GLMs cause the difficulties in the modelling of the complicated relationship between the response and predictor variables. Compared with statistical models, machine learning methods tend to have some advantages such as their ability to model non-linear relationships, remain robust despite missing data and outliers, reduce overfitting, etc. (Breiman 2001; Friedman and Meulman 2003).

REFERENCES

Basak, D., Pal, S., and Patranabis, D. C. 2007. Support vector regression. *Neural Information Processing-Letters and Reviews*, *11*(10), 203-224.

Breiman, L. 2001. Random forests. *Machine learning*, *45*(1), 5-32.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Friedman, J. H. and Meulman, J. J. 2003. Multiple additive regression trees with application in epidemiology. *Statistics in medicine*, *22*(9), 1365-1381.

Guisan, A., Weiss, S. B., and Weiss, A. D. 1999. GLM versus CCA spatial modeling of plant species distribution. *Plant Ecology*, *143*(1), 107-122.

Hastie, T. J. and Tibshirani, R. J. 1990. *Generalized additive models* (*Volume 43*: *Monographs on Statistics and Applied Probability*). CRC press, Chapman & Hall, London.

McCullagh, P. and Nelder, J. A. (ed.). 1989. *Generalized linear models* (*Vol. 37*). CRC press, Chapman & Hall, London, 513pp.

Moisen, G. G., Freeman, E. A., Blackard, J. A., Frescino, T. S., Zimmermann, N. E., & Edwards Jr, T. C. 2006. Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecological modelling*, *199*(2), 176-187.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Van Houwelingen, J. C. and Le Cessie, S. 1990. Predictive value of statistical models. *Statistics in medicine*, *9*(11), 1303-1325.

Yee, T. W. and Mitchell, N. D. 1991. Generalized additive models in plant ecology. *Journal of vegetation science*, *2*(5), 587-602.

Zaniewski, A. E., Lehmann, A. and Overton, J. M. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological modelling*, *157*(2-3), 261-280.

Table 1 Main features of the applied methods in this study and details of model settings in R.

Method	**Abbreviations**	**Main features**	**Function and settings in R**
Generalized linear models	GLM	The extension of linear regression model allowing various types of error distributions for response variable other than a normal distribution.	Function "glm" with family="gaussian"
Generalized additive models	GAM	The semi-parametric extension of GLM maintaining the additivity of the modelling approach, and allowing for non-linear functions in the predictor variables.	Function "gam" in package "gam" with family="gaussian"
Least absolute shrinkage and selection operator	Lasso	Allows for the shrinkage and selection of predictor variables by optimizing the loss function with 1-norm penalty term.	Function "cv.glmnet" in package "glmnet" with family="gaussian",nfolds = 10, type.measure = "mse"
Support vector machine	SVM	Minimizes the loss function with a tolerance function of the error.	Function "svm" in package "e1071" with type="eps-regression", kernel="radial",cost=1,gamma=5
Random forests	RF	Enhances many indentically distributed deep regression trees and averages the results of all trees to get a more accurate and stable prediction.	Function "randomForest" in package "randomForest" with ntree=1000,mtry =6
Boosted regression trees	BRT	Enhances the additive shallow trees to the residuals to make a strong learner.	Function "gbm" in package "gbm" with shrinkage = 0.005, distribution = 'gaussian', cv.folds = 5,n.trees = 5000

Table 2 Statistics of the predicted mean density of survey area provided by each method when CV=10%. True mean density on this area conditioned by RF is 20.189 g/m2. Values marked in red indicate the lowest value in each of the rows.

	MEAN		MEDIAN		RELATIVE BIAS				RMSE	
					mean		median			
Methods	*zz*	*tt*	*zz*	*tt*	*zz*	*tt*	*zz*	*tt*	*zz*	*tt*
naive estimator	26.058	26.202	25.512	26.533	29.1%	29.8%	26.4%	31.4%	6.319	6.315
GLM	7526.731	396.811	23.397	19.634	37181.7%	1865.5%	15.9%	-2.7%	27255.555	1568.199
GAM	30.818	30.61	20.877	20.462	52.6%	51.6%	3.4%	1.4%	27.153	51.152
Lasso	20.027	20.311	20.151	20.23	**-0.8%**	**0.6%**	-0.2%	**0.2%**	1.287	0.978
RF	23.09	20.614	20.429	19.946	14.4%	2.1%	1.2%	-1.2%	7.895	1.936
BRT	25.437	21.327	20.782	19.32	26.0%	5.6%	2.9%	-4.3%	12.678	4.745
SVM	21.029	21.521	20.197	21.777	4.2%	6.6%	**0.0%**	7.9%	2.285	1.806

Table 3 Statistics of the predicted mean density of survey area provided by each method when CV=30%. True mean density on this area conditioned by RF is 20.205 g/m2. Values marked in red indicate the lowest value in each of the rows.

	MEAN		MEDIAN		RELATIVE BIAS				RMSE	
					mean		median			
Methods	*zz*	*tt*	*zz*	*tt*	*zz*	*tt*	*zz*	*tt*	*zz*	*tt*
naive estimator	25.86	26.04	25.213	26.542	28.0%	28.9%	24.8%	31.4%	6.087	6.121
GLM	6475.887	247.302	21.823	19.866	31951.1%	1124.0%	8.0%	**-1.7%**	35625.613	1185.283
GAM	40.159	75.539	20.296	21.175	98.8%	273.9%	**0.5%**	4.8%	69.163	431.766
Lasso	19.124	19.448	18.967	19.433	-5.4%	-3.7%	-6.1%	-3.8%	1.65	1.144
RF	22.673	20.249	19.951	19.486	12.2%	**0.2%**	-1.3%	-3.6%	7.706	2.358
BRT	26.159	21.095	20.338	19.366	29.5%	4.4%	0.7%	-4.2%	14.609	4.411
SVM	20.05	20.564	19.267	20.616	**-0.8%**	1.8%	-4.6%	2.0%	1.986	1.14

Table 4 Statistics of the predicted mean density of survey area provided by each method when CV=50%. True mean density on this area conditioned by RF is 20.189 g/m2. Values marked in red indicate the lowest value in each of the rows.

	MEAN		MEDIAN		RELATIVE BIAS				RMSE	
					mean		median			
Methods	*zz*	*tt*	*zz*	*tt*	*zz*	*tt*	*zz*	*tt*	*zz*	*tt*
naive estimator	25.997	26.116	25.391	26.47	28.8%	29.4%	25.8%	31.1%	6.228	6.209
GLM	6151.296	241.28	22.805	19.229	30368.5%	1095.1%	13.0%	-4.8%	29727.476	995.336
GAM	79.398	68.465	19.238	19.22	293.3%	239.1%	**-4.7%**	-4.8%	240.383	312.594
Lasso	17.729	18.151	17.584	18.133	-12.2%	-10.1%	-12.9%	-10.2%	2.714	2.237
RF	20.736	18.869	18.473	18.381	**2.7%**	-6.5%	-8.5%	-9.0%	6.773	2.686
BRT	24.064	19.607	19.082	18.008	19.2%	**-2.9%**	-5.5%	-10.8%	13.193	5.061
SVM	18.809	19.305	18.072	19.345	-6.8%	-4.4%	-10.5%	**-4.2%**	2.312	1.326

Table 5 Statistics of the predicted mean density of survey area provided by each method when CV=10%. True mean density on this area conditioned by BRT is 26.654 g/m2. Values marked in red indicate the lowest value in each of the rows.

	MEAN		MEDIAN		RELATIVE BIAS				RMSE	
					mean		median			
Methods	*zz*	*tt*	*zz*	*tt*	*zz*	*tt*	*zz*	*tt*	*zz*	*tt*
naive estimator	30.811	30.738	32.544	32.865	15.6%	15.3%	22.1%	23.3%	6.603	6.064
GLM	1759540.3	152036.41	22.897	23.224	6601321.5%	570308.3%	-14.1%	-12.9%	8675475.12	803483.42
GAM	25.608	22.443	21.782	21.207	-3.9%	-15.8%	-18.3%	-20.4%	12.203	6.832
Lasso	22.152	22.833	22.822	23.386	-16.9%	-14.3%	-14.4%	-12.3%	4.959	4.239
RF	26.919	26.055	26.21	26.354	**1.0%**	**-2.2%**	-1.7%	**-1.1%**	4.133	3.18
BRT	26.242	25.062	26.375	24.995	-1.5%	-6.0%	**-1.0%**	-6.2%	2.625	3.044
SVM	23.119	23.585	22.955	24.324	-13.3%	-11.5%	-13.9%	-8.7%	4.646	3.815

Table 6 Statistics of the predicted mean density of survey area provided by each method when CV=30%. True mean density on this area conditioned by BRT is 26.613 g/m2. Values marked in red indicate the lowest value in each of the rows.

	MEAN		MEDIAN		RELATIVE BIAS				RMSE	
					mean		median			
Methods	*zz*	*tt*	*zz*	*tt*	*zz*	*tt*	*zz*	*tt*	*zz*	*tt*
naive estimator	29.811	29.803	30.301	31.447	12.0%	12.0%	13.9%	18.2%	6.13	5.668
GLM	826536.63	32606.274	22.764	22.137	3105687.9%	122421.1%	-14.5%	-16.8%	6471059.03	232061.47
GAM	25.7	21.963	20.559	20.293	**-3.4%**	-17.5%	-22.7%	-23.7%	14.077	7.597
Lasso	20.819	21.502	21.428	21.742	-21.8%	-19.2%	-19.5%	-18.3%	6.129	5.46
RF	25.636	24.693	24.621	24.965	-3.7%	**-7.2%**	-7.5%	**-6.2%**	4.048	3.74
BRT	25.06	24.035	24.786	24.011	-5.8%	-9.7%	**-6.9%**	-9.8%	3.153	3.69
SVM	22	22.502	21.539	22.996	-17.3%	-15.4%	-19.1%	-13.6%	5.508	4.724

Table 7 Statistics of the predicted mean density of survey area provided by each method when CV=50%. True mean density on this area conditioned by BRT is 26.567 g/m2. Values marked in red indicate the lowest value in each of the rows.

	MEAN		MEDIAN		RELATIVE BIAS				RMSE	
					mean		median			
Methods	*zz*	*tt*	*zz*	*tt*	*zz*	*tt*	*zz*	*tt*	*zz*	*tt*
naive estimator	30.614	30.366	30.294	31	15.2%	14.3%	14.0%	18.6%	6.701	6.066
GLM	437789.243	63010.099	21.355	21	1647750.3%	237071.7%	-19.6%	-19.2%	2395768.067	468598.522
GAM	23.011	22.153	20.231	19	-13.4%	-16.6%	-23.9%	-27.3%	11.829	9.434
Lasso	19.64	20.161	20.095	21	-26.1%	-24.1%	-24.4%	-21.4%	7.182	6.651
RF	23.864	23.286	23.781	23	**-10.2%**	**-12.3%**	**-10.5%**	**-11.8%**	4.457	4.43
BRT	23.516	22.622	23.487	22	-11.5%	-14.9%	-11.6%	-15.8%	4.194	4.686
SVM	21.124	21.647	20.826	22	-20.5%	-18.5%	-21.6%	-16.9%	6.113	5.394

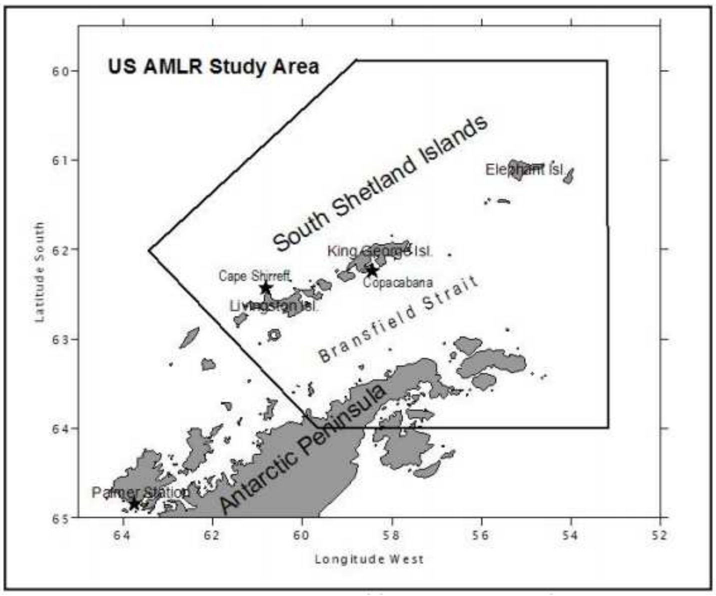

Figure 1 Locations of U.S. AMLR Field Stations: Cape Shirreff, Livingston Island; Admiralty Bay (Copacabana), King George Island.

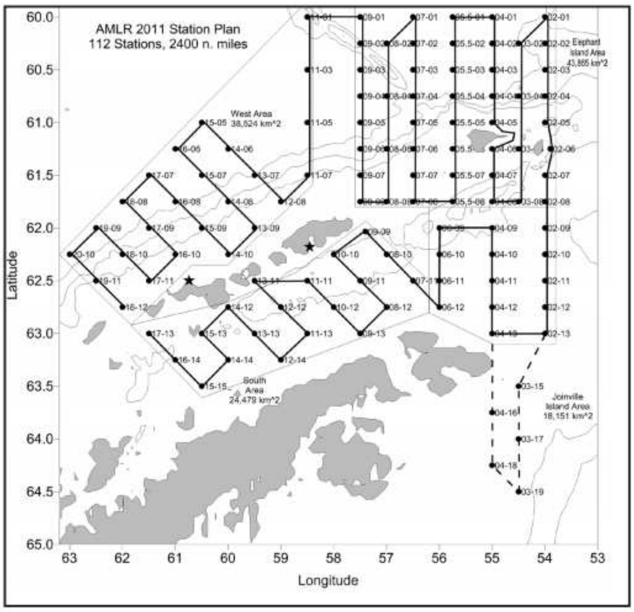

Figure 2 The survey design of AMLR 2010/11 (Leg I/Survey A) in the vicinity of the South Shetland Islands, Elephant Island, and the Antarctic Peninsula. Black dots indicate locations of planned oceanographic/biological sampling stations and heavy lines indicate planned transects between stations.

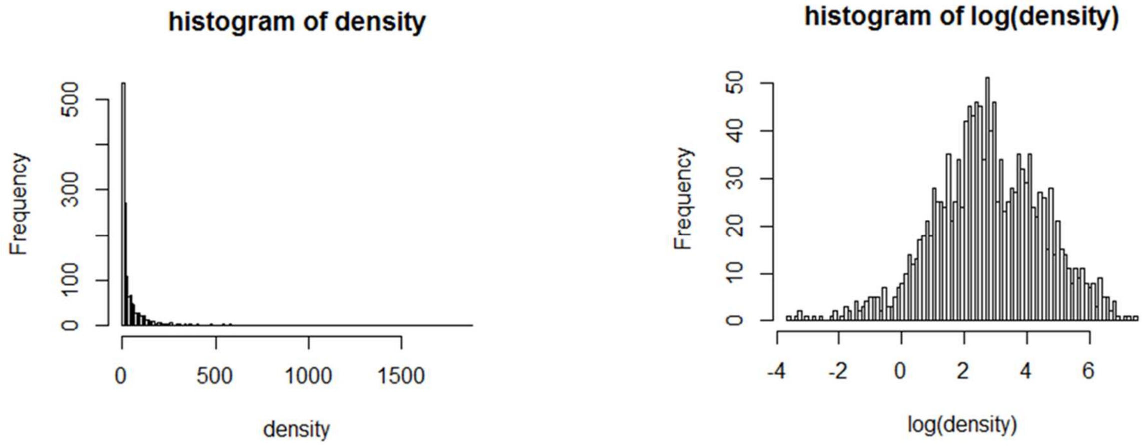

Figure 3 Histograms of observed density (left panel) and observed density in log scale (right panel).

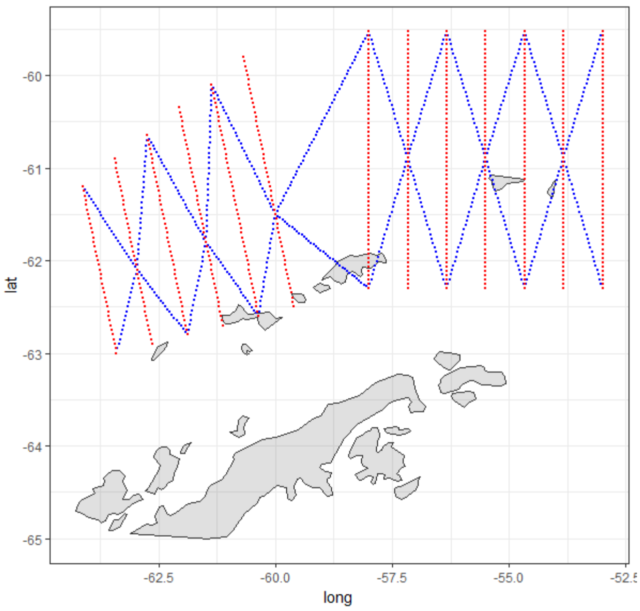

Figure 4 Simulation of the survey designs of the study area. The blue dashed line indicates the effective survey routes of the tooth design, with the red dashed lines indicating the effective zigzag survey routes. In total, 689 stations are set in the zigzag design, and 696 stations are set for the tooth design.

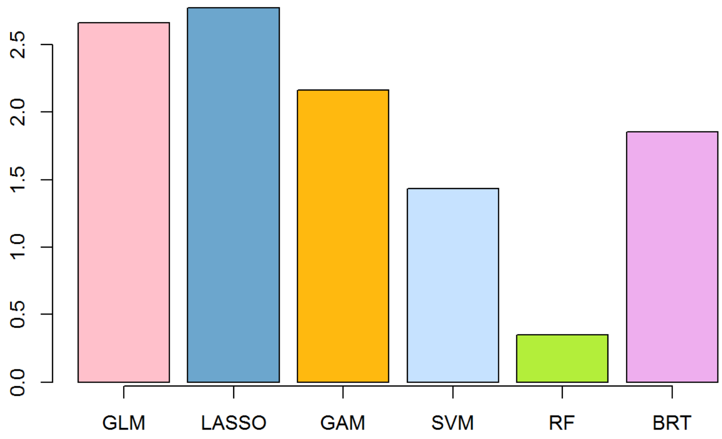

Figure 5 A histogram illustrating the RSS of the six models for fitting with total data set. All the calculations were conducted in the log scale.

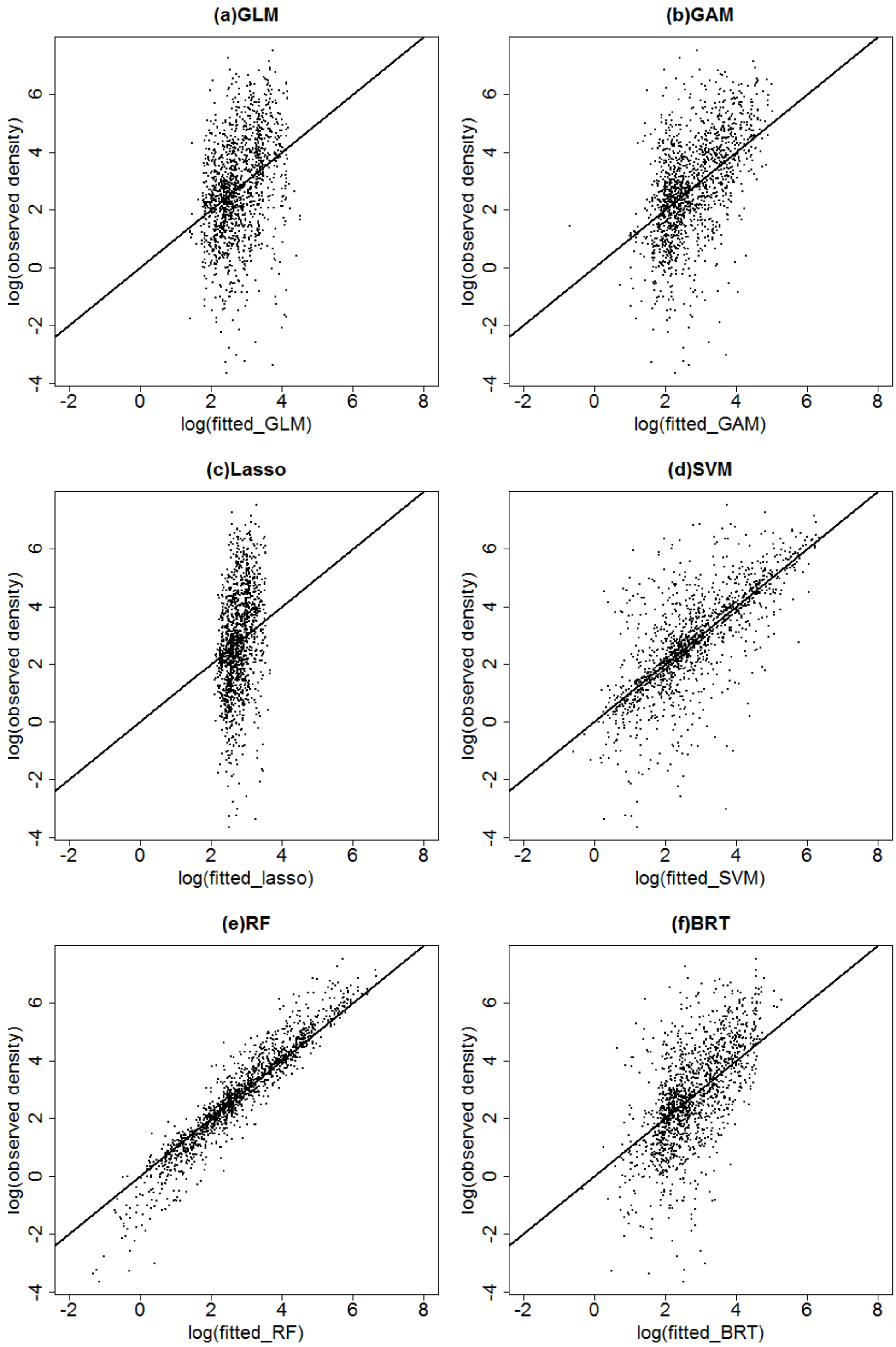

Figure 6 Scatter plots of each model for the comparison of fitting results and density observations. Horizontal axis indicates the fitting results, and vertical axis indicates the density observations. The density values for both axes are in log scale.

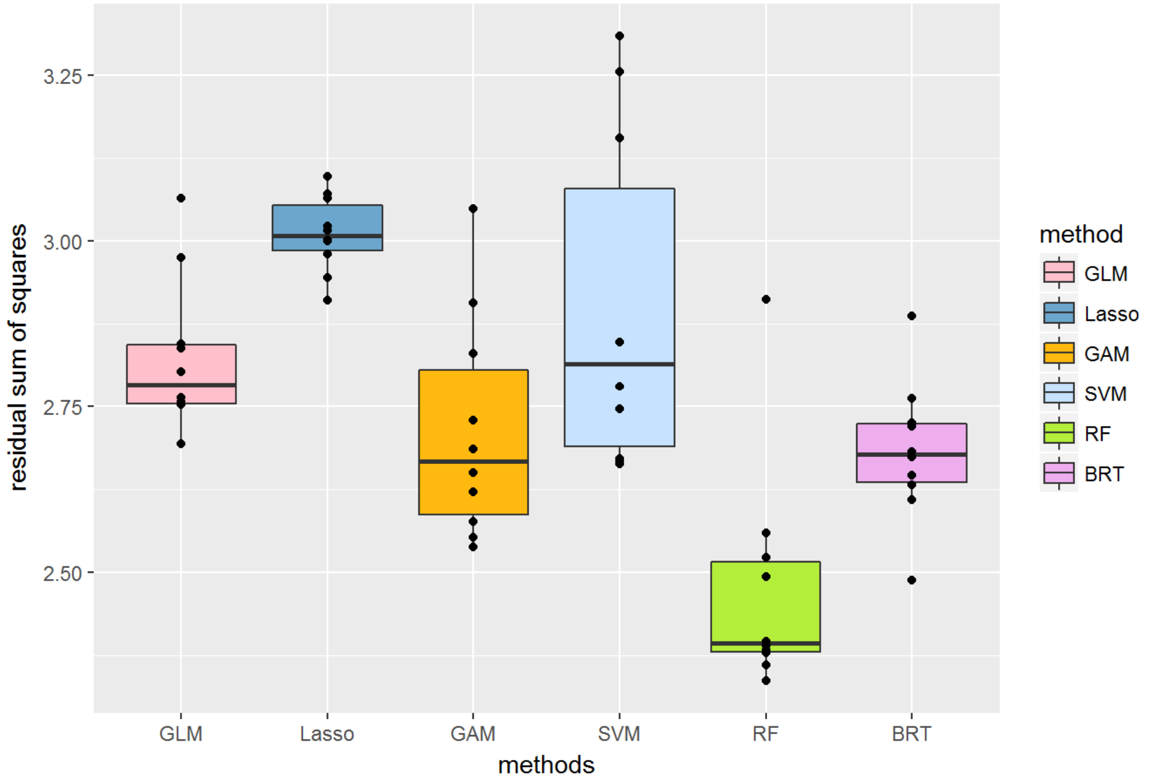

Figure 7 A boxplot illustrating the RSS of the six models for 10-fold cross-validation. The boxes show median and 1st and 3rd quartile values. All the calculations are conducted in the log scale.

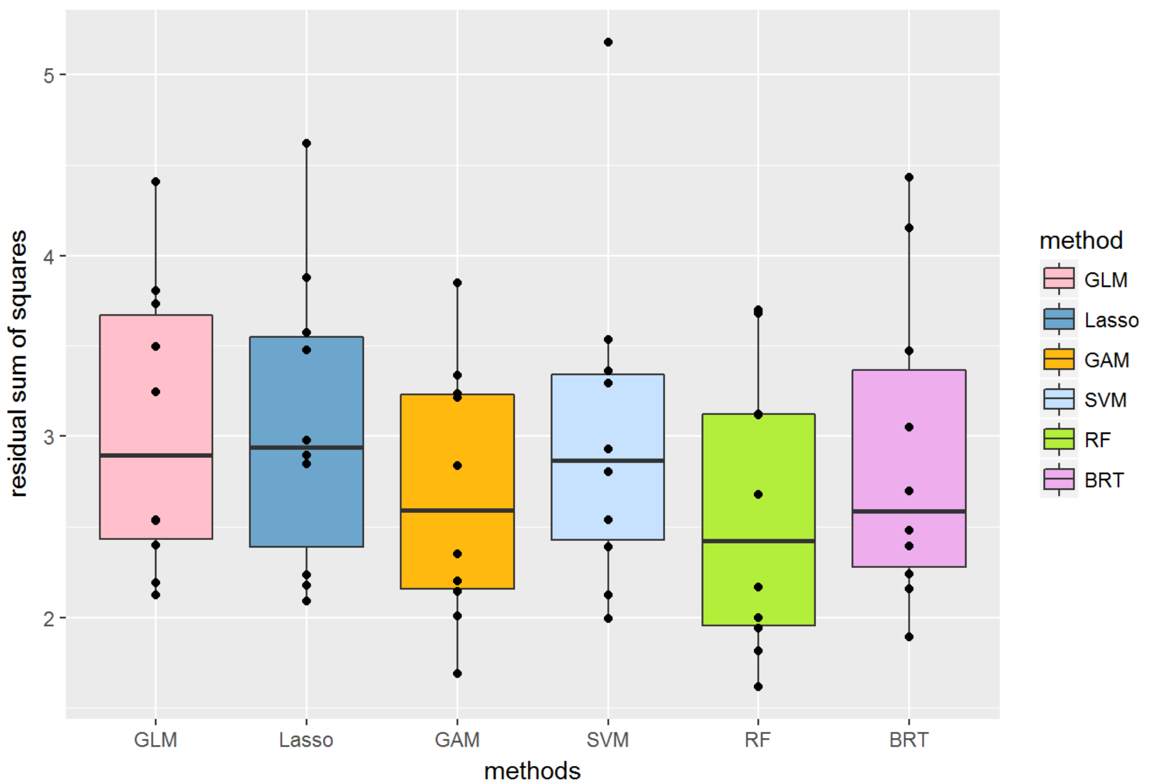

Figure 8 A boxplot illustrating the RSS of the six models for spatial cross-validation. The boxes show median and 1st and 3rd quartile values. All the calculations are conducted in the log scale.

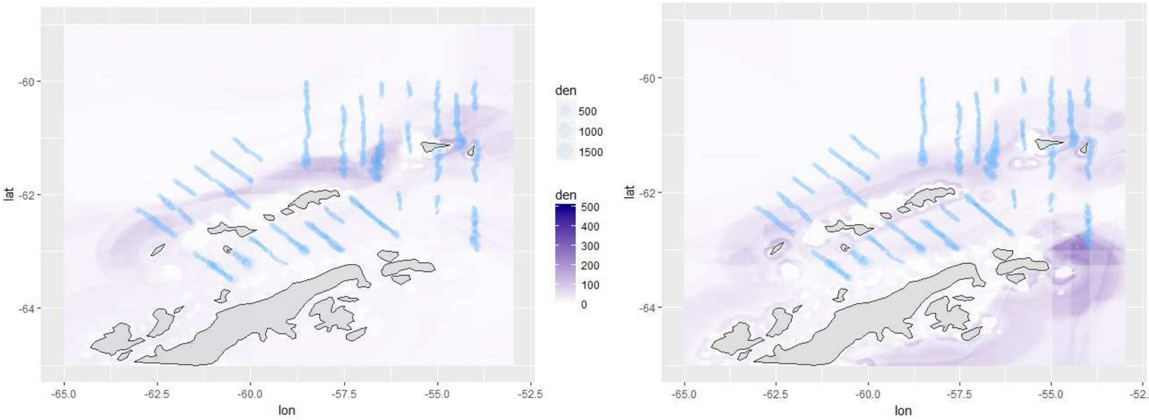

Figure 9 Krill density distribution surfaces conditioned by random forests (left panel) and boosted regression trees (right panel) in the study region. Density observations are colored in blue, with the higher density plotted in larger size. The predictions are colored in purple, with the deeper color indicating higher predicted density.

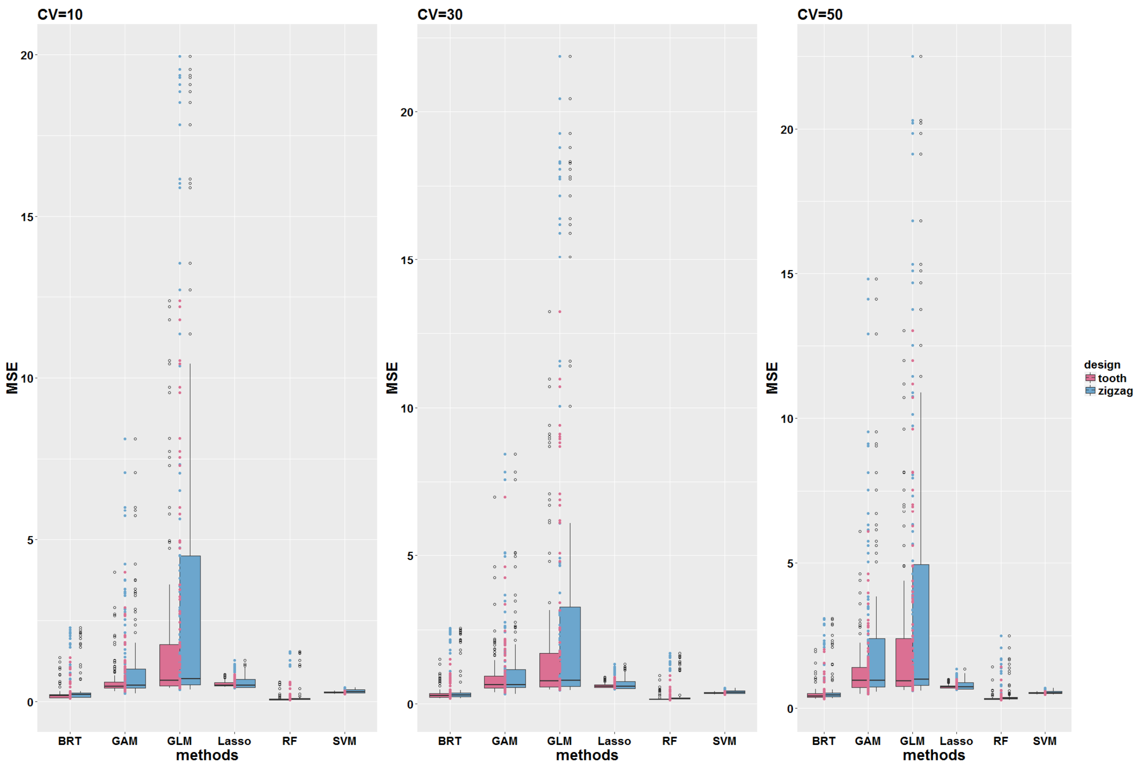

Figure 10 A box plot illustrating the MSE of the predicted density distribution for RF surface provided by each model under the tooth (colored in pink) and zigzag (colored in blue) design with CV=10%, 30% and 50% respectively.

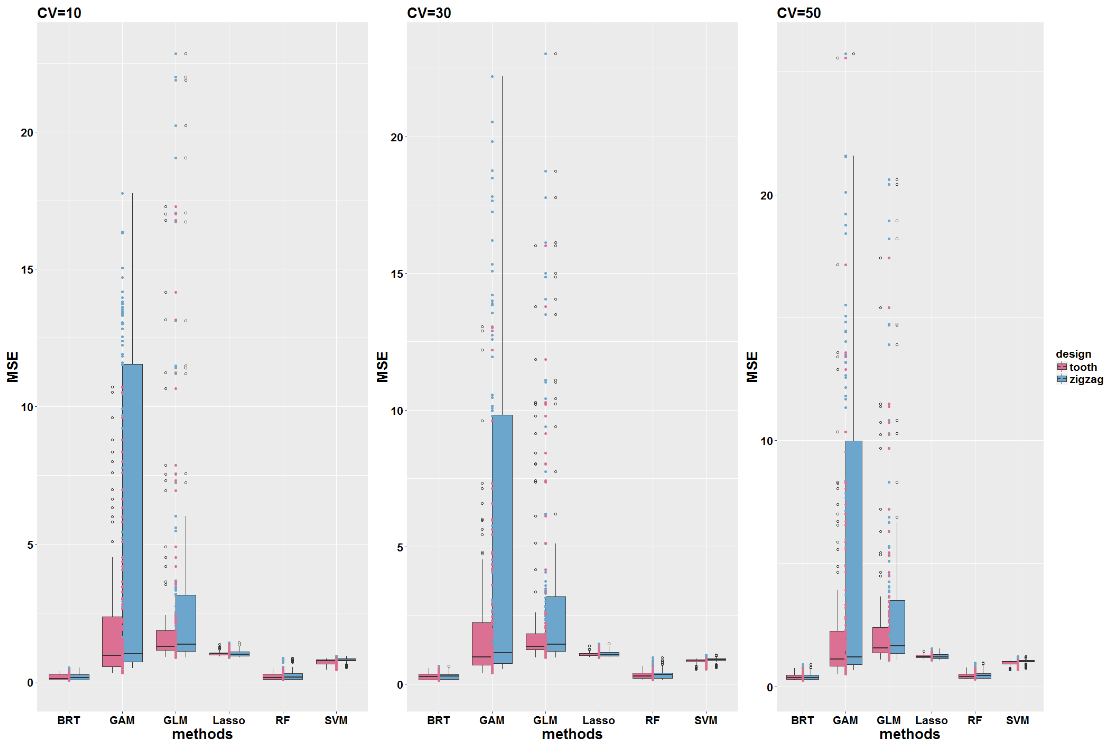

Figure 11 A box plot illustrating the MSE of the predicted density distribution for BRT surface provided by each model under the tooth (colored in pink) and zigzag (colored in blue) design with CV=10%, 30% and 50% respectively.

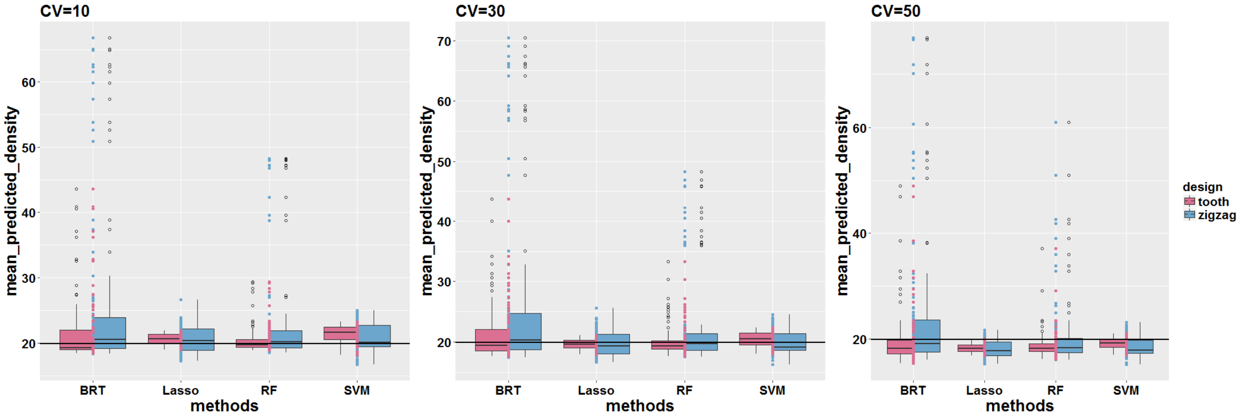

Figure 12 A box plot illustrating the predicted mean density value of the predicted density distribution for RF surface provided by each model under the tooth (colored in pink) and zigzag (colored in blue) design with CV=10%, 30% and 50% respectively. The horizontal lines indicate the true mean density value conditioned by RF. Only the results of four methods, BRT, Lasso, RF and SVM are depicted here, while the results of GLM and GAM are out of the range of this figure.

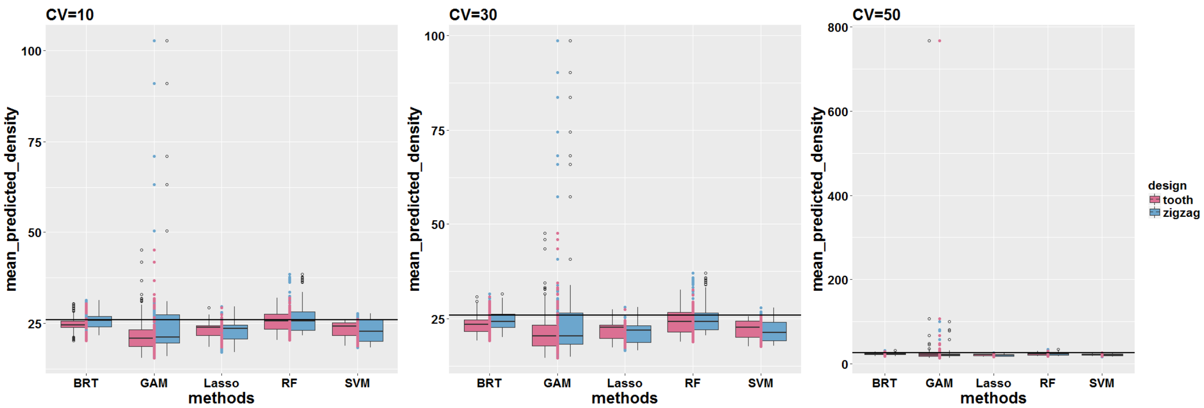

Figure 13 A box plot illustrating the predicted mean density value of the predicted density distribution for BRT surface provided by each model under the tooth (colored in pink) and zigzag (colored in blue) design with CV=10%, 30% and 50% respectively. The horizontal lines indicate the true mean density value conditioned by BRT. Only the results of five methods, BRT, GAM, Lasso, RF and SVM are depicted here, while the results of GLM is out of the range of this figure.

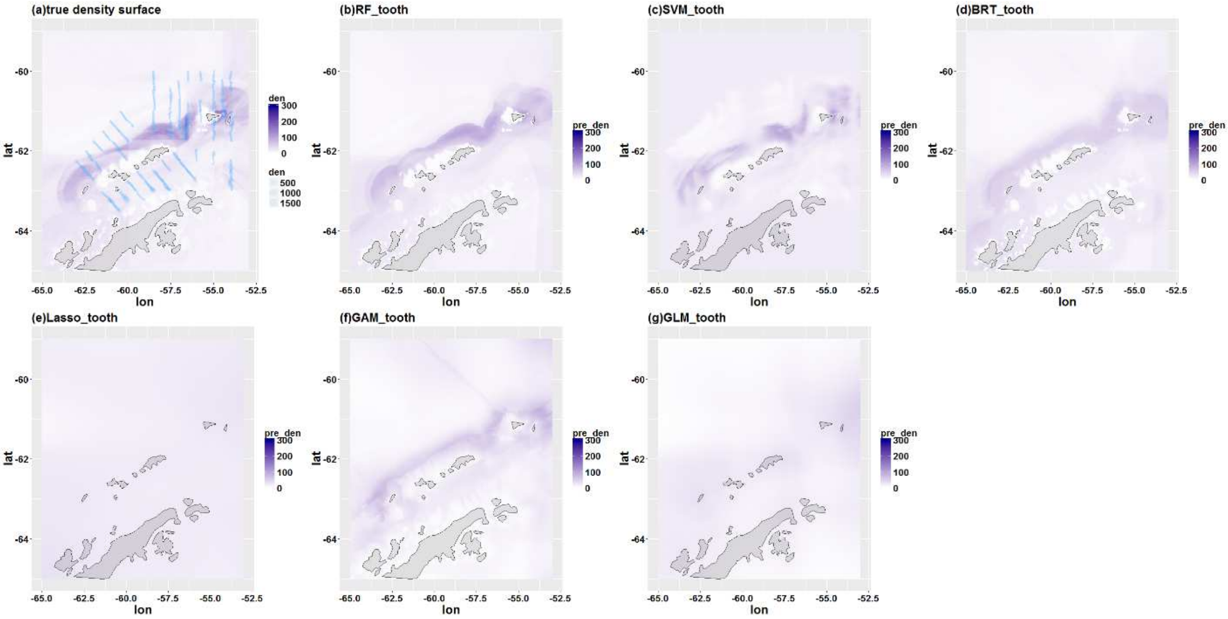

Figure 14 Predicted spatial density distribution of krill for RF surface provided by six models with the tooth survey design, CV=10%. (a) conditioned true density distribution surface by random forests(RF) (b) predicted density distribution by random forests(RF) (c) predicted density distribution by support vector machines(SVM) (d) predicted density distribution by boosted regression trees (BRT) (e) predicted density distribution by least absolute shrinkage and selection operator (Lasso) (f) predicted density distribution by generalized additive models (GAM) (g) predicted density distribution by generalized linear models (GLM). The scale of the legend ranges from 0 to 300 grams per square meter.

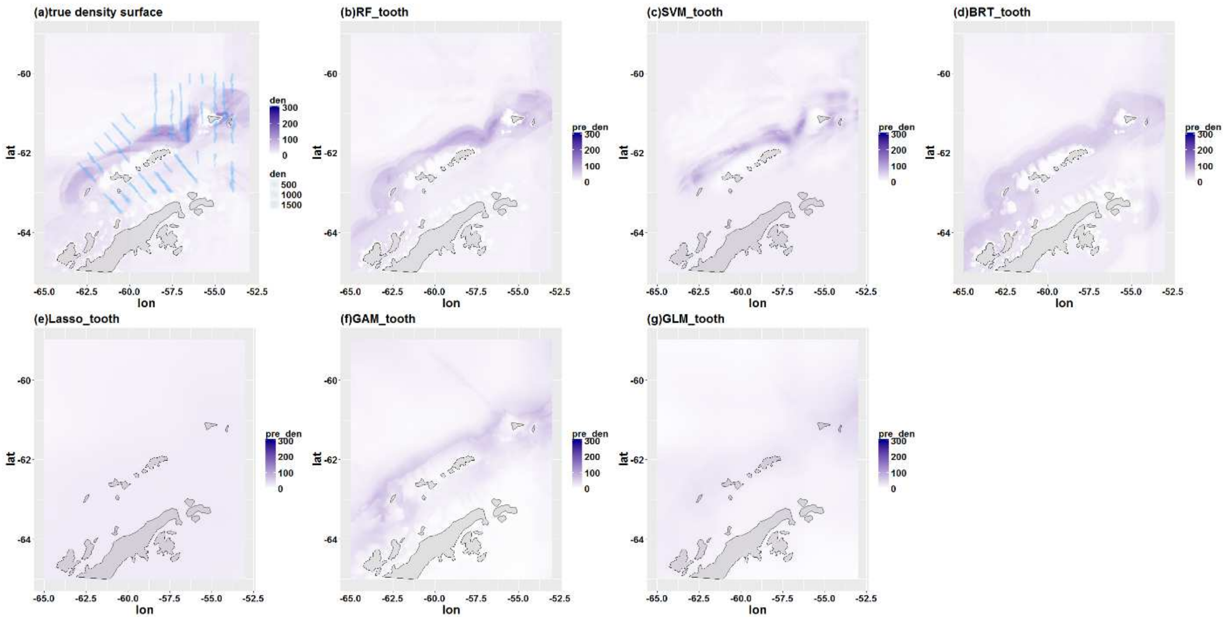

Figure 15 Predicted spatial density distribution of krill for RF surface provided by six models with the tooth survey design, CV=30%.

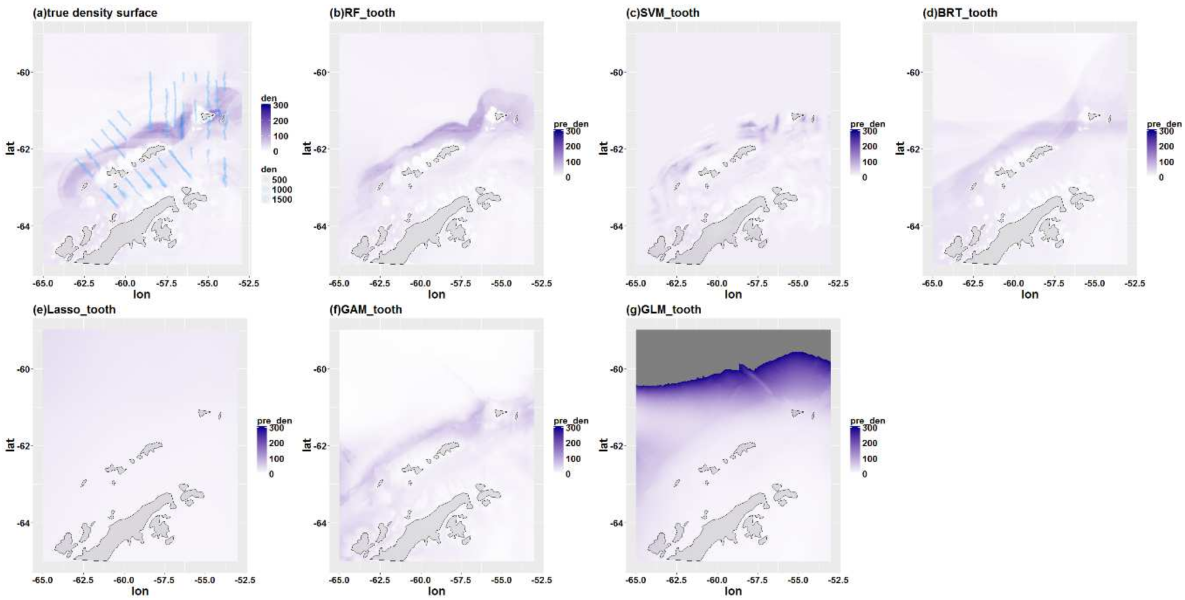

Figure 16 Predicted spatial density distribution of krill for RF surface provided by six models with the tooth survey design, CV=50%.

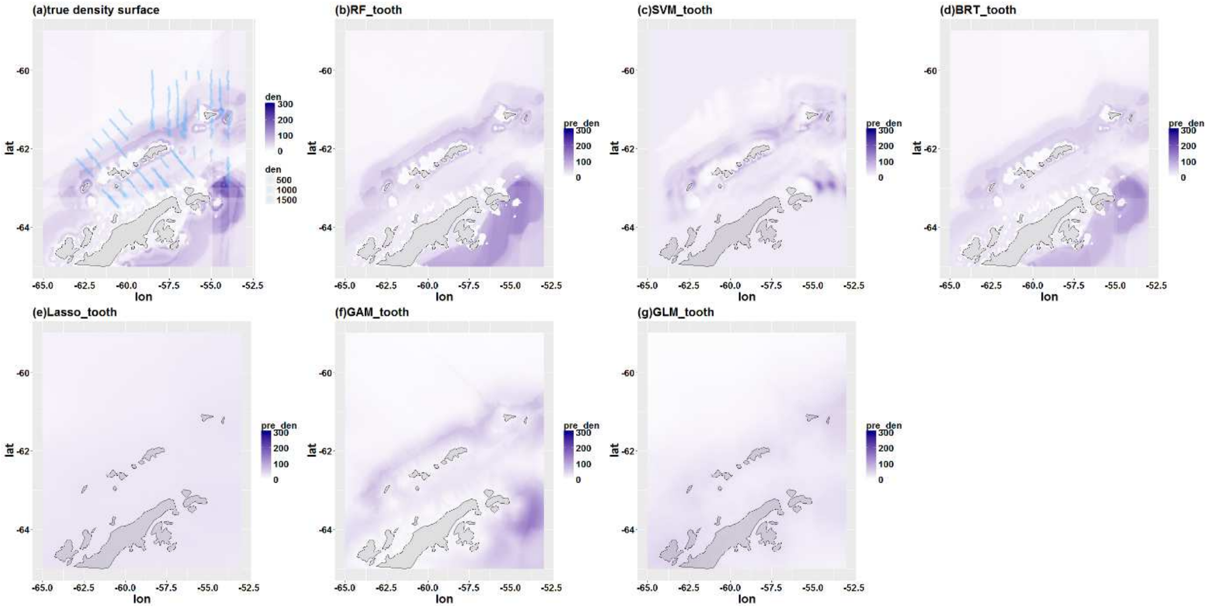

Figure 17 Predicted spatial density distribution of krill for BRT surface provided by six models with the tooth survey design, CV=10%. (a) conditioned true density distribution surface by boosted regression trees(BRT) (b) predicted density distribution by random forests(RF) (c) predicted density distribution by support vector machines(SVM) (d) predicted density distribution by boosted regression trees (BRT) (e) predicted density distribution by least absolute shrinkage and selection operator (Lasso) (f) predicted density distribution by generalized additive models (GAM) (g) predicted density distribution by generalized linear models (GLM). The scale of the legend ranges from 0 to 300 grams per square meter.

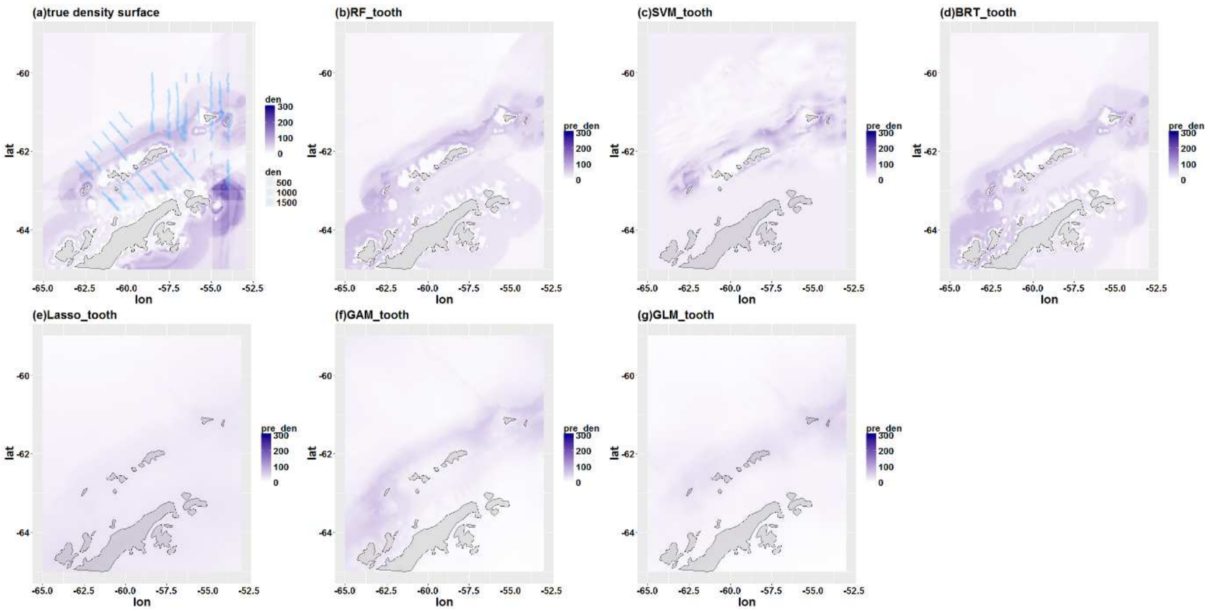

Figure 18 Predicted spatial density distribution of krill for BRT surface provided by six models with the tooth survey design, CV=30%.

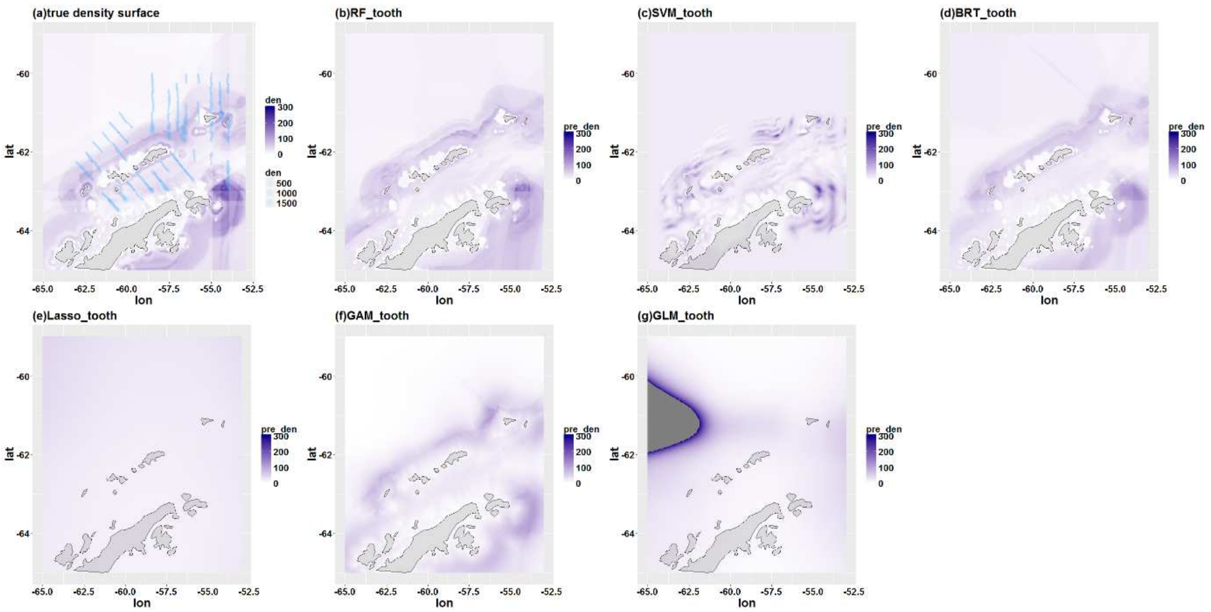

Figure 19 Predicted spatial density distribution of krill for BRT surface provided by six models with the tooth survey design, CV=50%.