Filling in Sea Surface Temperatures

YanSong

Introduction

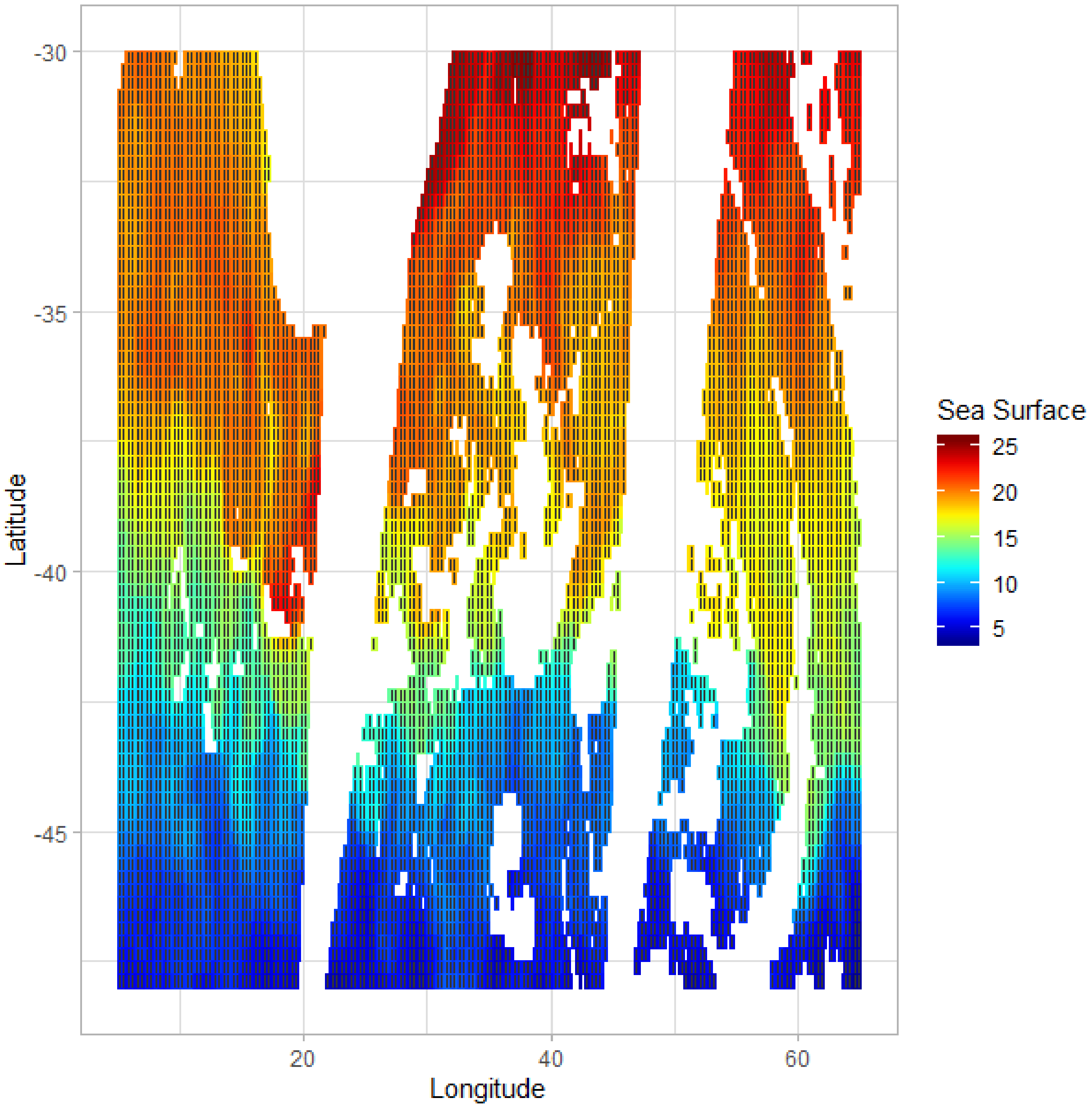
West-blowing trade winds in the Indian Ocean push warm surface waters against the eastern coast of Africa. These waters move south along the coastline, eventually spilling out along the boundary of the Indian and Atlantic Oceans. This jet of warm water, known as the Agulhas Current, collides with the cold, west to east flowing Antarctic Circumpolar Current, producing a dynamic series of meanders and eddies as the two waters mix.

The data SST.mat file contains sea surface temperature data collected by satellite for the Agulhas and surrounding areas off the coast of South Africa from January 1 to November 26, 2004, a period of 331 days. The data contains a lot of missing values which caused by land, satellite's orbital clipping and cloud cover.

In this project, we aim to predict the missing values present in the Day 10 data with Gaussian geostatistical model. However, the model need $O(n^3)$ numerical operations to estimate parameters and make a prediction. Calculating with full data is time consuming. Thus, we consider a data reduction. Here we employ three subsampling methods, including random, deep and wide, and MaxProLHD methods. We compare their parameter estimation and prediction performance. Then we choose the best one and fill in the missing value.

Subsampling Methods

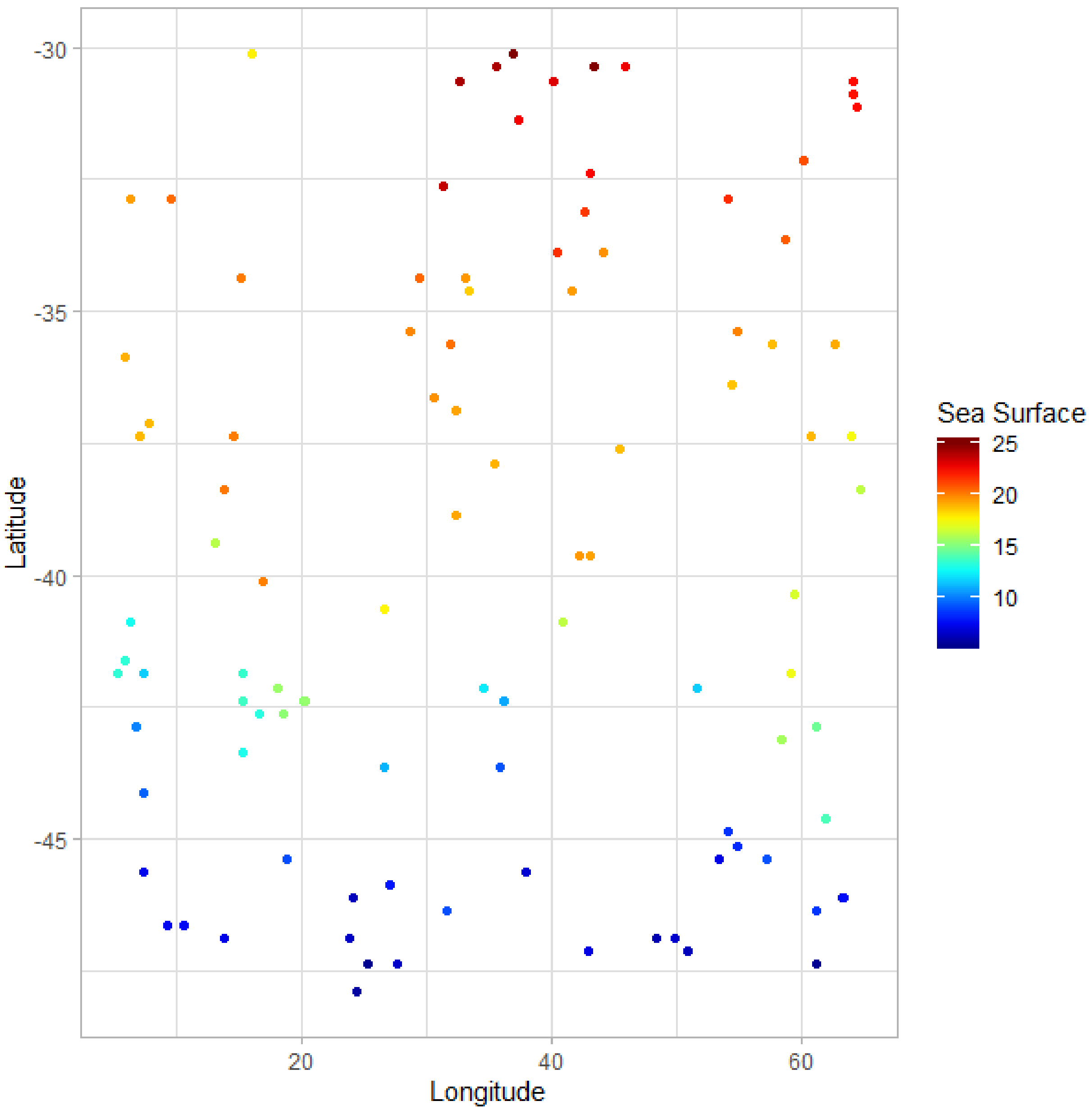
In this section, we introduce three subsampling methods. Suppose that we need m subsamples. Figure 1 depicts the sea surface temperatures recorded on day 10.



> Figure 1. Sea surface temperatures on day 10.

Random

Choose m subsamples randomly from full data. Random subsampling method is the easiest one. Figure 2 shows 100 random subsamples. We can see that the subsamples scatter irregularly on the whole region.



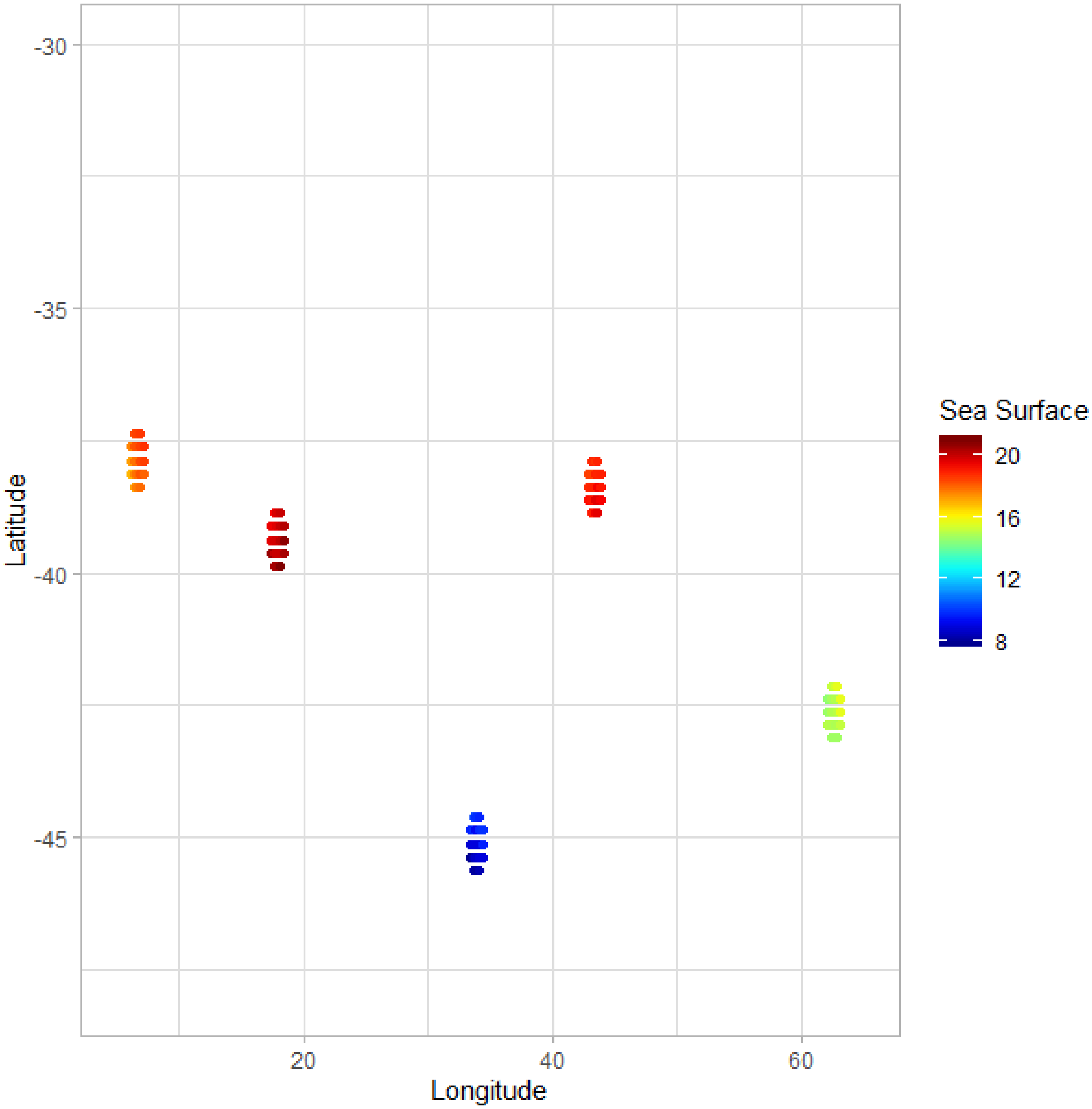
> Figure 2. 100 subsamples selected Randomly.

Deep and Wide (DaW)

Step 1: Choose five subsamples randomly, called center points.

Step 2: For each center point, choose its $\frac{m}{5} - 1$ nearest subsamples.

Figure 3 displays 100 DaW subsamples. The subsamples gather in five groups, that is, "deep". The five groups distribute in the whole region, which is "wide". As shown in Figure 3, the DaW subsamples can not cover the region well. However, in spatial statistics, the "deep" structure can help us estimate the covariance parameters well.



> Figure 3. 100 subsamples selected by DaW.

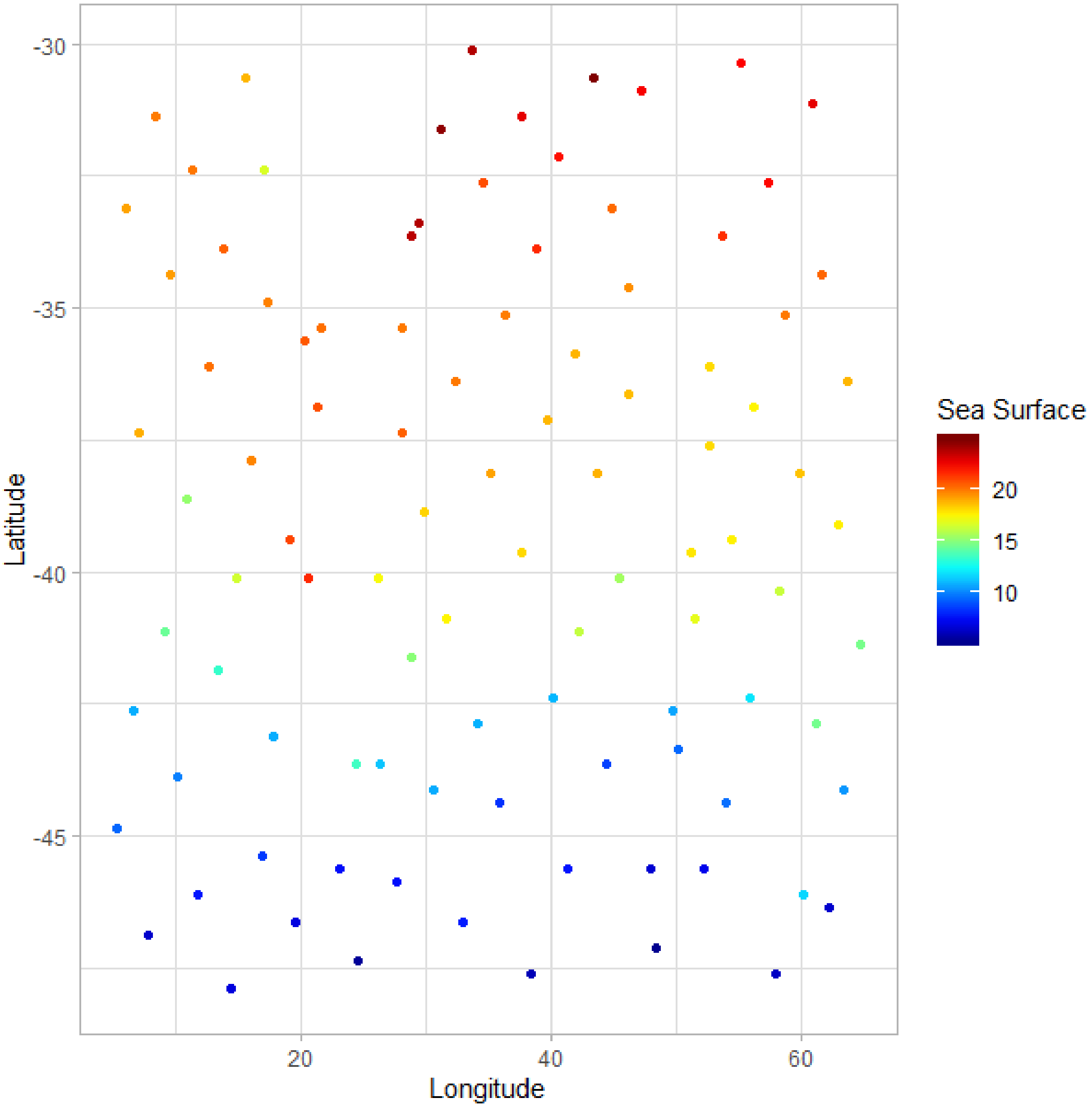
Maximum Projection Design (MaxPro)

MaxPro is a kind of Latin Hypercube design proposed by Joseph et al. (2015). It has space-filling property and ensures good projections to all subspaces of the factors. That is, the design points can cover all subspace as uniformly as possible. The MaxPro design points can be obtained from R package *MaxPro*. Then subsamples can be chosen by following algorithm.

Step 1: Generete m design points by *MaxPro*.

Step 2: Select the nearest neighbor for each design point from full data.

Figure 4 illustrates 100 MaxPro subsamples. Compared with Random subsamples, MaxPro subsamples scatter more uniformly. Compared with DaW subsamples, MaxPro subsamples cover the region better but have no “deep” structure.



> Figure 4. 100 subsamples selected by MaxPro.

Performance of various subsamples

Model selection

Recall that we have selected Gaussian linear geostatistical model to fit the data. The model is given by

$$y(\mathbf{s}_i) = \mathbf{x}_i^T \beta + w(\mathbf{s}_i) + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d}}{\sim} \mathrm{N}(0, \tau^2)$$

where $y(\mathbf{s}_i)$ is the observation at location $\mathbf{s}_i$, $\mathbf{x}_i = (1, \mathbf{s}_i^T)^T$, $\beta = (\beta_1, \beta_2, \beta_3)$ is a 3-dimentional vector of unknown parameters, $\{w(\mathbf{s}_i)\}$ is a spatial Gaussian process with $\mathrm{E}(w(\mathbf{s}_i)) = 0$, $\mathrm{Var}(w(\mathbf{s}_i)) = \sigma^2$, and

$\mathrm{Corr}(w(\mathbf{s}_i), w(\mathbf{s}_j)) = \exp(-\frac{\|\mathbf{s}_i - \mathbf{s}_j\|}{\phi})$. Under this model, we have

$$\mathbf{y} \sim \mathrm{N}(X\beta, \Sigma(\phi, \sigma^2)),$$

where $\Sigma(\phi, \sigma^2)$ is not a sparse matrix. R package *geoR* provides corresponding functions of parameter estimation and prediction.

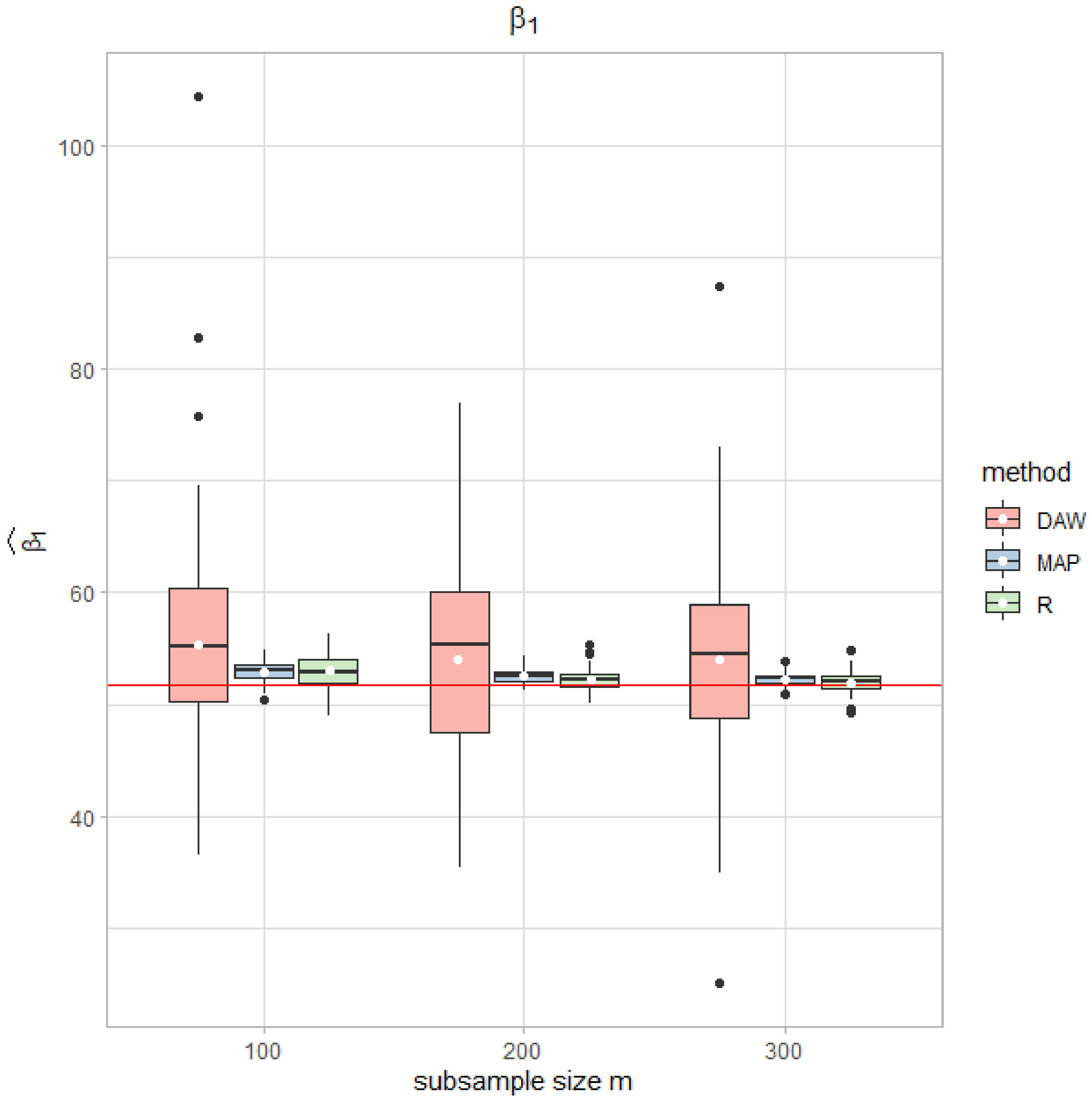
Comparison of three subsampling methods

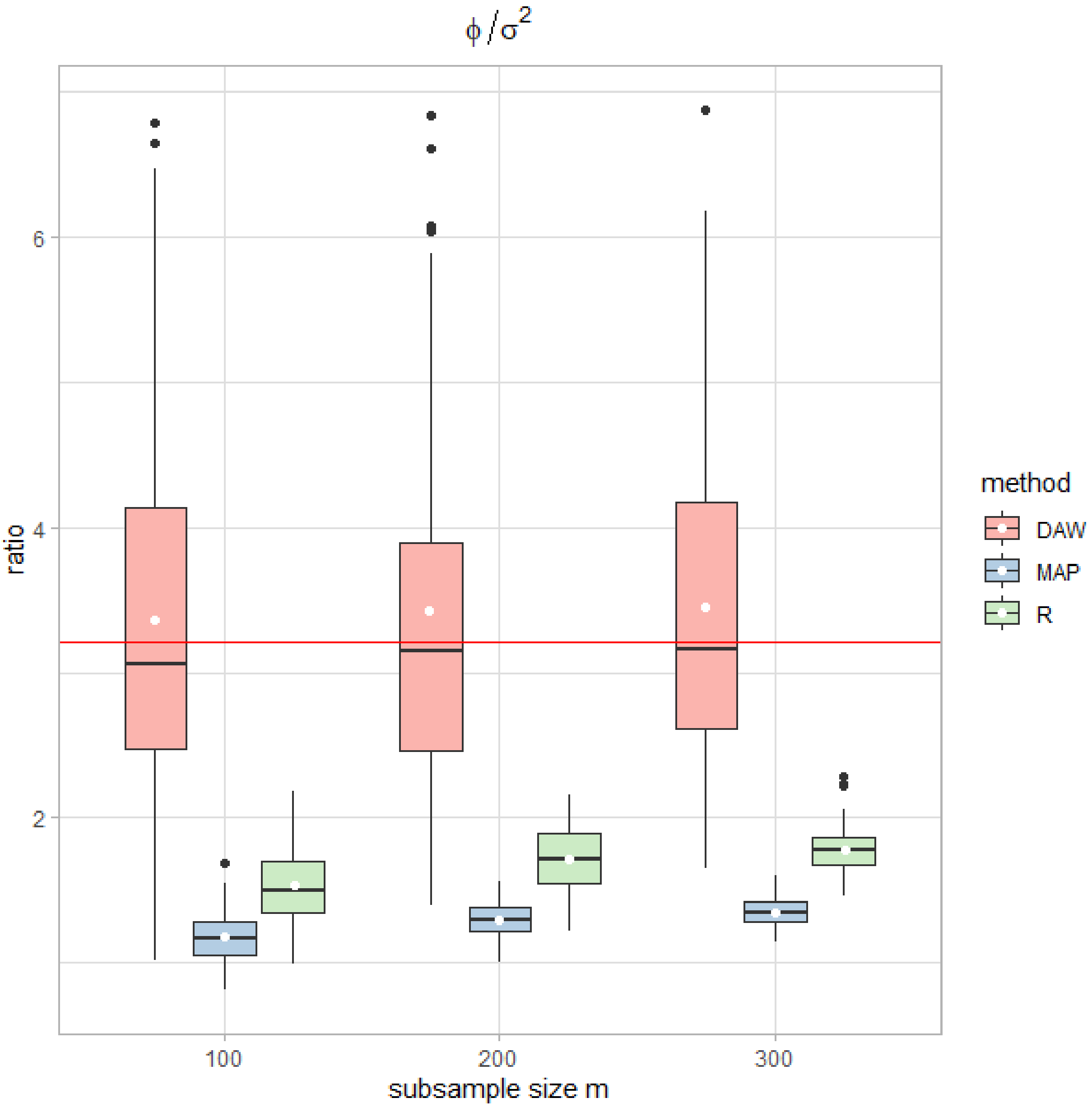
In this section, we illustrate the performance of the three subsampling methods under various subsample sizes m. We choose the best one by comparison.

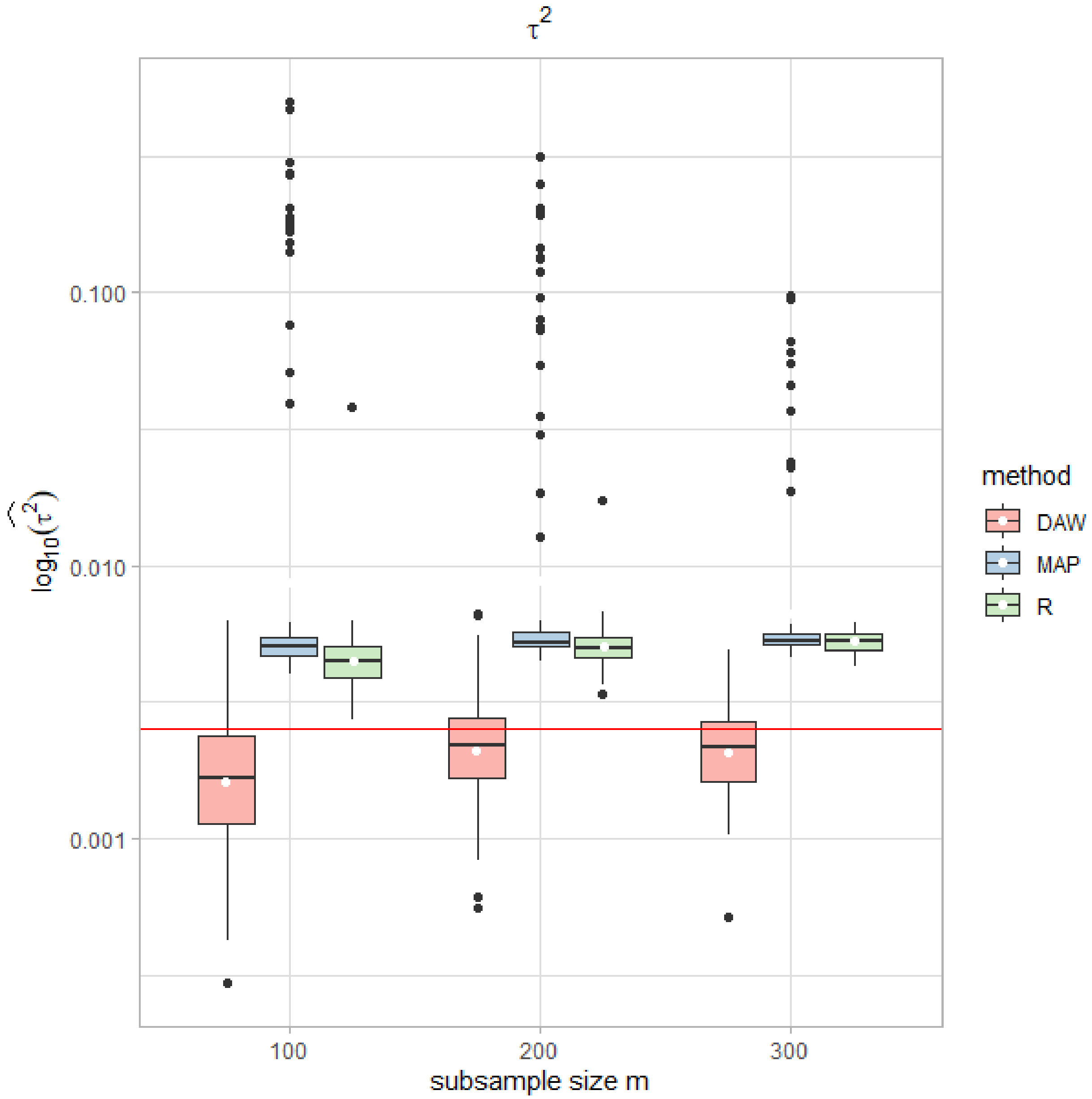
In our analysis, we first devided the full data into two parts, a random subset of 4000 observations as the training set and the remaining 6945 observations as the test data. We treated the training data as "full data", whose parameter estimation and prediction results can be regarded as standards. Then, using various subsampling methods, we chose m subsamples from the training data. The subsample size m was set to be 100, 200 and 300 so that we could observe how these methods perform over m. We utilized subsamples to estimate the parameters and predict the test data. We replicated the above process 100 times and summarized the outcomes as follows.

Parameter estimation

For each combination of m and subsampling method, we have 100 results. We summarized them into Figure 5. The top left panel of Figure 5 illustrates the estimation of β_1. We can see that MaxPro subsampling method had the best performance, especially when m is not large. DaW behaved worst, since it had a larger bias and variance. β_2 and β_3 had similar performances to β_1, so their results are not shown. The top right panel of Figure 5 depicts the estimation of covariance parameters. DaW had a much lower bias because its deep structure is helpful for covariance parameters' estimation. The variance of DaW was large, since the random selection of only five center points brought much uncertainty. The bottom left panel shows the estimation of nugget, which was similar to that of covariance parameters. Based on the assumption that the results of training set is true value, MSE of each parameter can be calculated. The bottom right panel shows the total MSEs of these parameters. We can see the total MSsE of MaxPro and DaW were close to each other and were much smaller than DaW.


β_1
100
80
60
40
$\hat{\beta}_1$
100
200
300
subsample size m
method
DAW
MAP
R


ϕ/σ^2
ratio
6
4
2
100
200
300
subsample size m
method
DAW
MAP
R


τ^2
$\log_{10}(\widehat{\tau^2})$
0.100
0.010
0.001
100
200
300
subsample size m
method
DAW
MAP
R

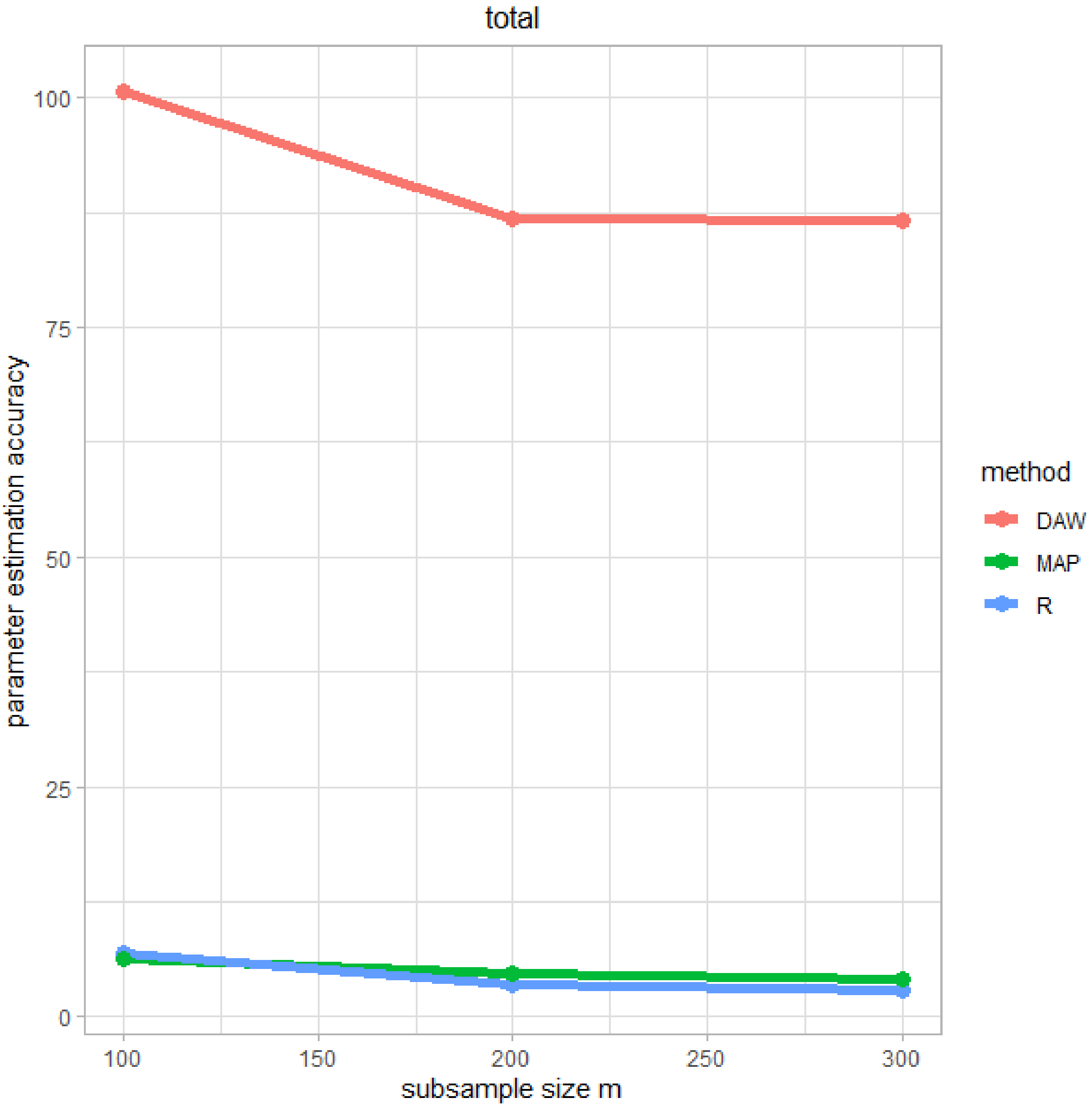


Figure 5. Performance of three subsampling methods in parameter estimation. The red line represents

Prediction accuracy

Using subsamples $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$, we can obtain the prediction of test data, denoted by $\hat{\mathbf{y}}$. Then we use mean squre prediction error (mspe), that is, $\frac{1}{m}(\hat{\mathbf{y}} - \mathbf{y})^T(\hat{\mathbf{y}} - \mathbf{y})$ to measure the prediction accuracy of subsamples. For each combination of m and method, we have 100 mspes. From Figure 6, we can see that MaxPro outperformed DaW and Random in term of prediction.

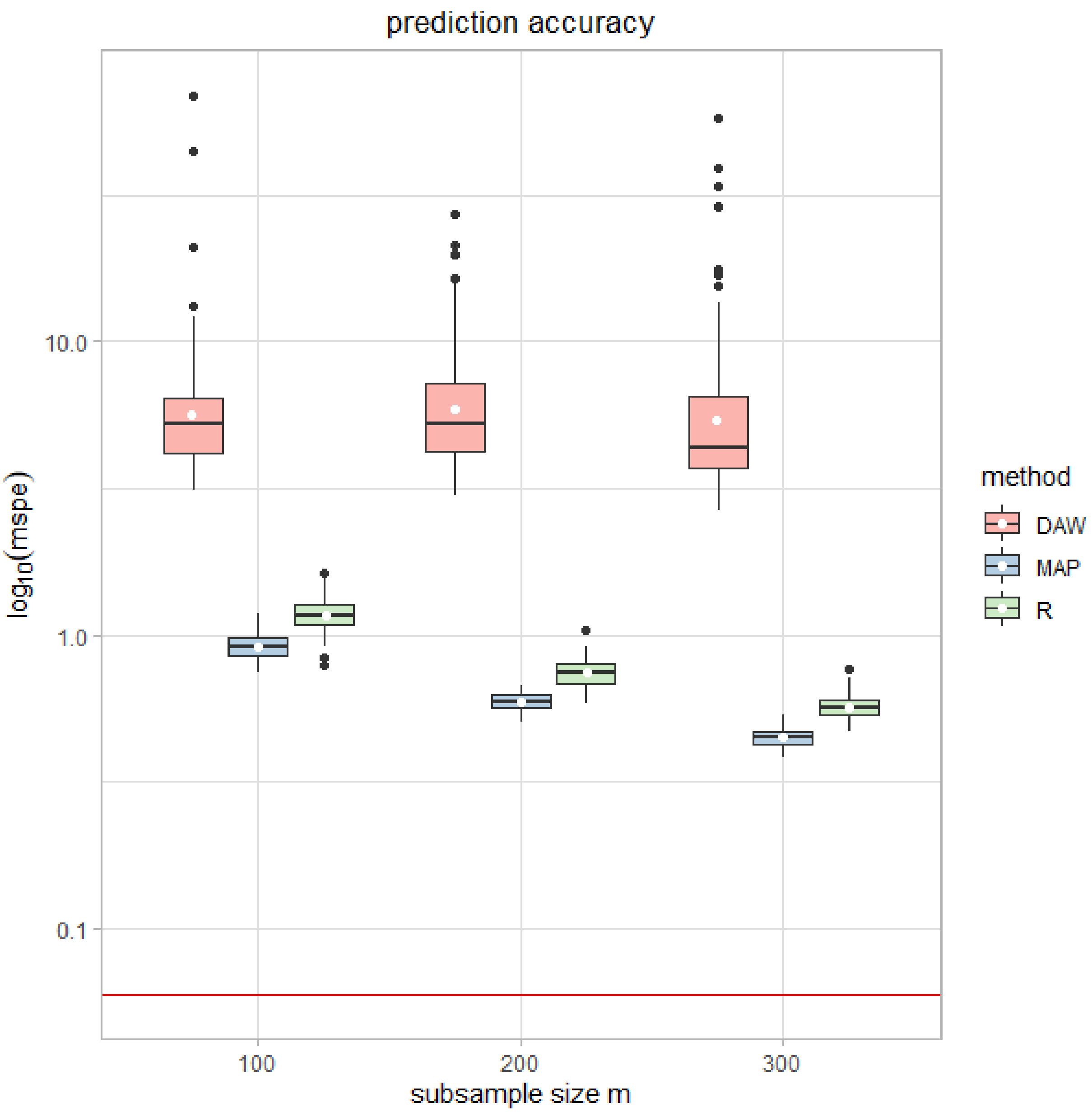


Figure 6. Performance of three subsampling methods in prediction. The red line represents results o

Fill in Missing Values

Based on the discussion above, MaxPro subsampling method was employed. We selected 300 MaxPro subsamples and used them to predict the missing values caused by cloud cover and satellite's orbital, as shown in Figure 7.

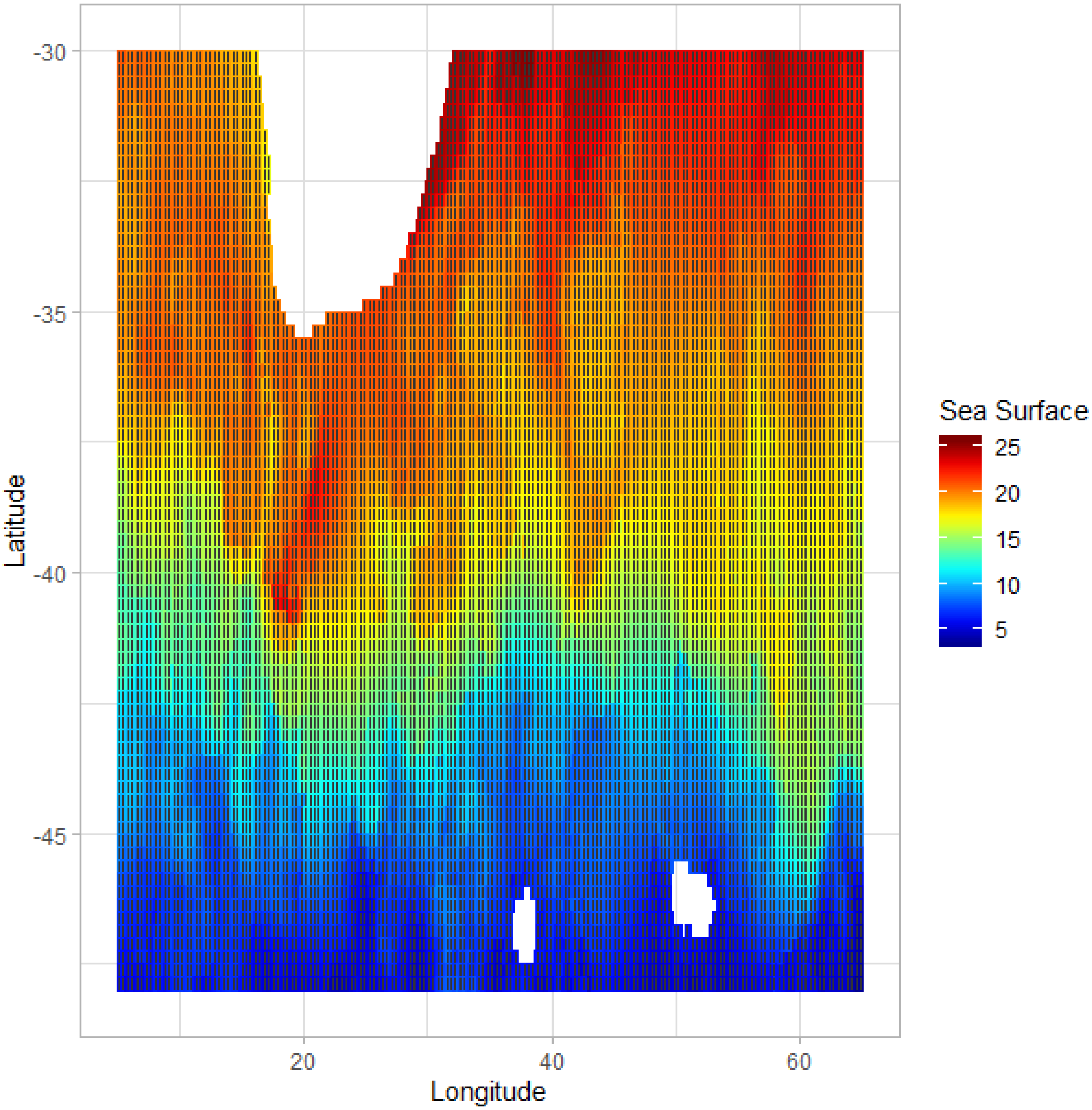


Figure 7. Sea surface temperatures on Day 10, with missing values estimaed by 300 MaxPro subsamples

Discussion

In this project, we introduced three subsampling methods, which are Random, DaW and MaxPro. By comparing their performances in parameter estimation and prediction, we finally chose MaxPro subsampling method to fill in the missing values in sea surface temperatures data on Day 10. MaxPro is a kind of space-filling design. Meng et al. (2020) selected MaxPro subsamples and used them to fit spline models. For our spatial data, MaxPro subsamples are better at estimating trend parameters and prediction. However, generating MaxPro design points is time consuming, which is a limitation of MaxPro subsampling method. When we estimate covariance parameters, DaW seems to have a smaller bias. The deep structure is benifit to estimate covariance parameters. Barbian and Assuncao (2017) have proposed an ensemble method based on DaW so that the variance can be reduced. When we care about the estimation of covariance parameters, we can employ ensemble DaW. Random method is the easiest method. From Figure 5 and Figure 6, we can see that results of Random were close to that of MaxPro. Therefore, when we want to fill in missing values roughly and quickly, we can choose Random methods. The selection of subsampling methods based on data and our target.

Referfence

E. B. S. Joseph, V. Roshan; Gul. Maximum projection designs for computer experiments. Biometrika, 102, 06 2015.

https://academic.oup.com/biomet/article/102/2/371/246859

C. Meng, X. Zhang, J. Zhang, W. Zhong, and P. Ma. More efficient approximation of smoothingsplines via space-filling basis selection, 2020.

https://arxiv.org/abs/2003.10609

M. H. Barbian and R. M. Assuncao. Spatial subsemble estimator for large geostatistical. Spatial Statistics, 22:68-88, 2017.

https://www.sciencedirect.com/science/article/abs/pii/S2211675316300999