Final Report

Yan Song

ISBD

May 7, 2020

Overview

1. Introduction
2. Subsampling Methods
3. Subsampling Methods Comparison
4. Fill in Missing Values
5. Discussion

Data Information

- Sea surface temperature data collected by satellite.
- Agulhas and surrounding areas off the coast of South Africa
- January 1 to November 26, 2004, a period of 331 days.
- A lot of missing values which were caused by land, satellite's orbital clipping and cloud cover.

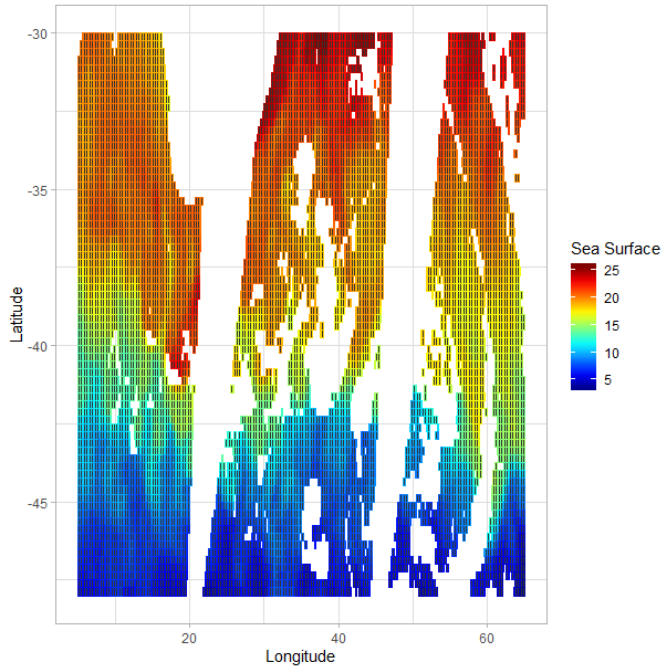

Figure: Sea surface temperature on Day 10.

Goal

- **Goal**: Fill in the missing values of Day 10 data.
- **Model**: Gaussian linear geostatistical model

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

$$\mathbf{y} \sim \mathrm{N}(X\beta, \Sigma(\theta)),$$

where the (i,j)th element of $\Sigma(\theta)$ is $\sigma^2 \exp\{-\frac{\|\mathbf{s}_i - \mathbf{s}_j\|}{\phi}\} + \tau^2 \mathbf{1}(\mathbf{s}_i = \mathbf{s}_j)$

- **Limitation**: Gaussian geostatistical model need $O(n^3)$ numerical operations to estimate parameters and make prediction. It is time consuming.

Solution

- **Solution**: Consider a data reduction.
- **Following section**:
  - Introduce three subsampling methods: Random, Deep and Wide, and MaxProLHD.
  - Compare the performance of various methods in parameter estimation and prediction.
  - Choose a subsampling method and use it to fill in the missing value.

Random

Choose m subsamples randomly from full data.

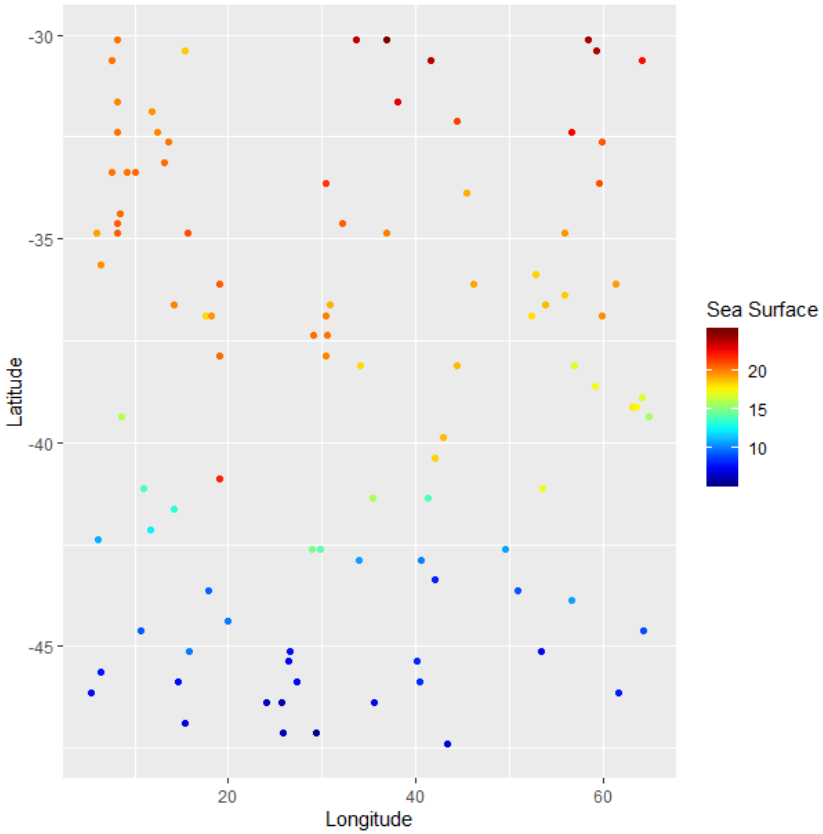

Figure: 100 Random subsamples

Deep and Wide (DaW)

- Step 1: Choose five subsamples randomly, called center points (Wide).
- Step 2: For each center point, choose its $\frac{m}{5} - 1$ nearest subsamples (Deep).

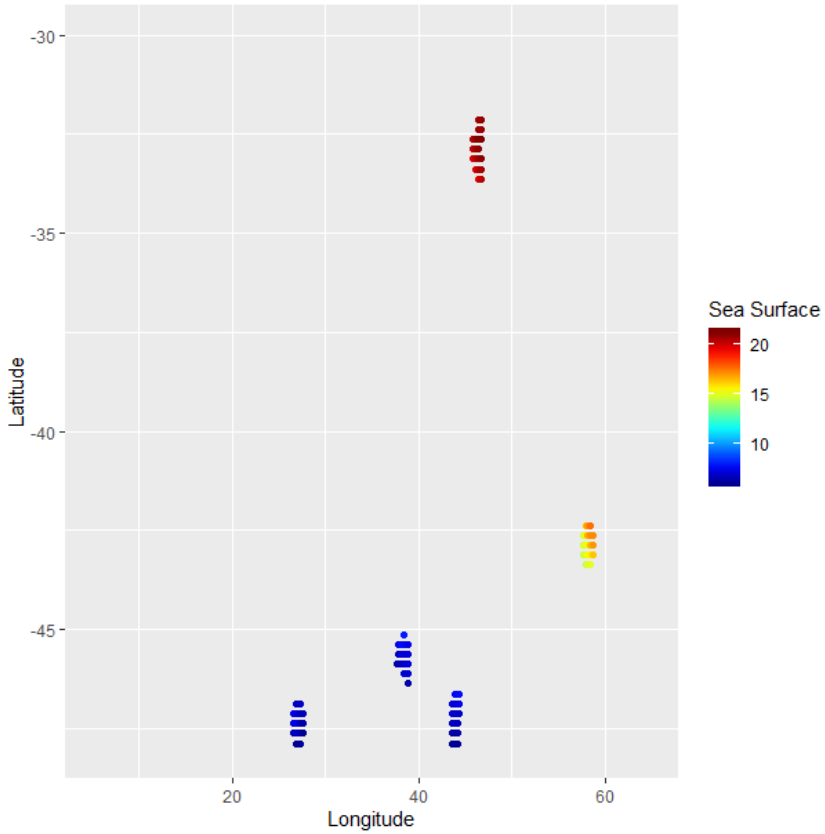

Figure: 100 DaW subsamples

Maximum Projection Design (MaxPro)

- Step 1: Generete m MaxPro design points by R package *MaxPro*.
- Step 2: Select the nearest neighbor for each design point from full data.

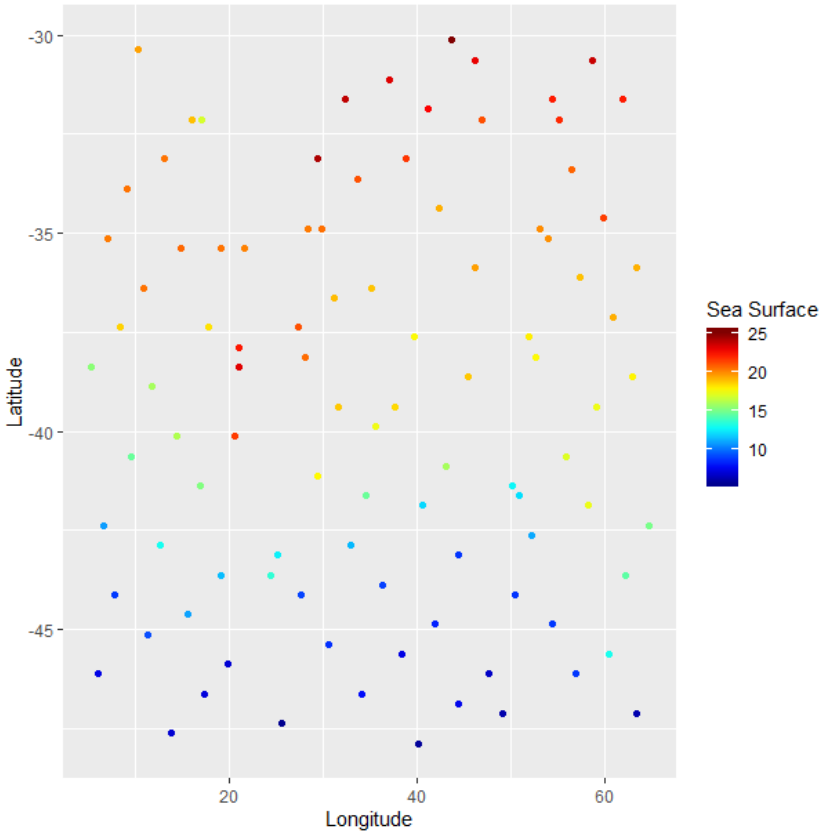

Figure: 100 MaxPro subsamples

Agorithm

The algorithm is summarized as follows.

- Step 1: Divide 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.
- Step 2: Estimate the parameters and predict the test data with training set, or "full data". Treat the results as standards.
- Step 3: Choose m subsamples from "full data" by various subsampling methods, where m=100, 200 and 300. Utilize subsamples to estimate the parameters and predict the test data.
- Step 4: Repeat step 3 100 times.

Parameter Estimation: Trend Parameters

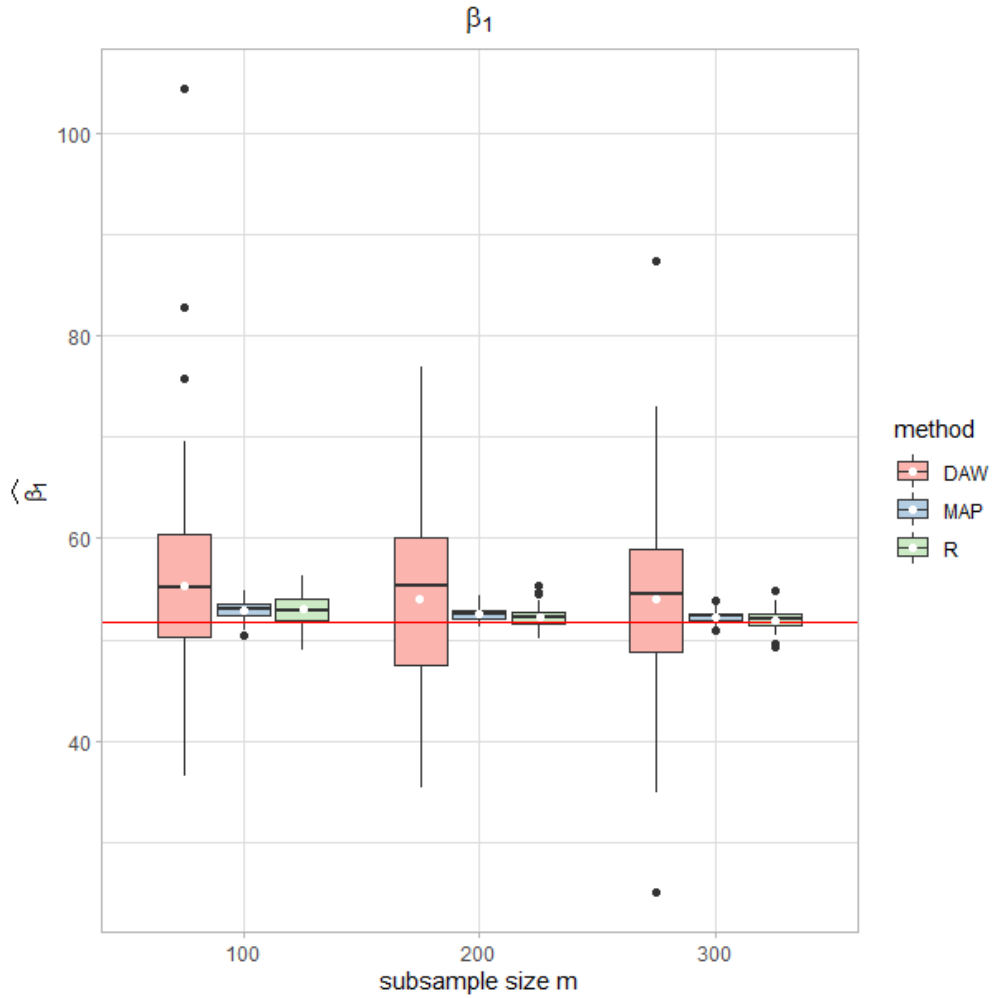

Figure: Boxplot of $\hat{\beta}_1$ estimated by various subsampling methods under m=100,200 and 300.

Parameter Estimation: Covariance Parameters

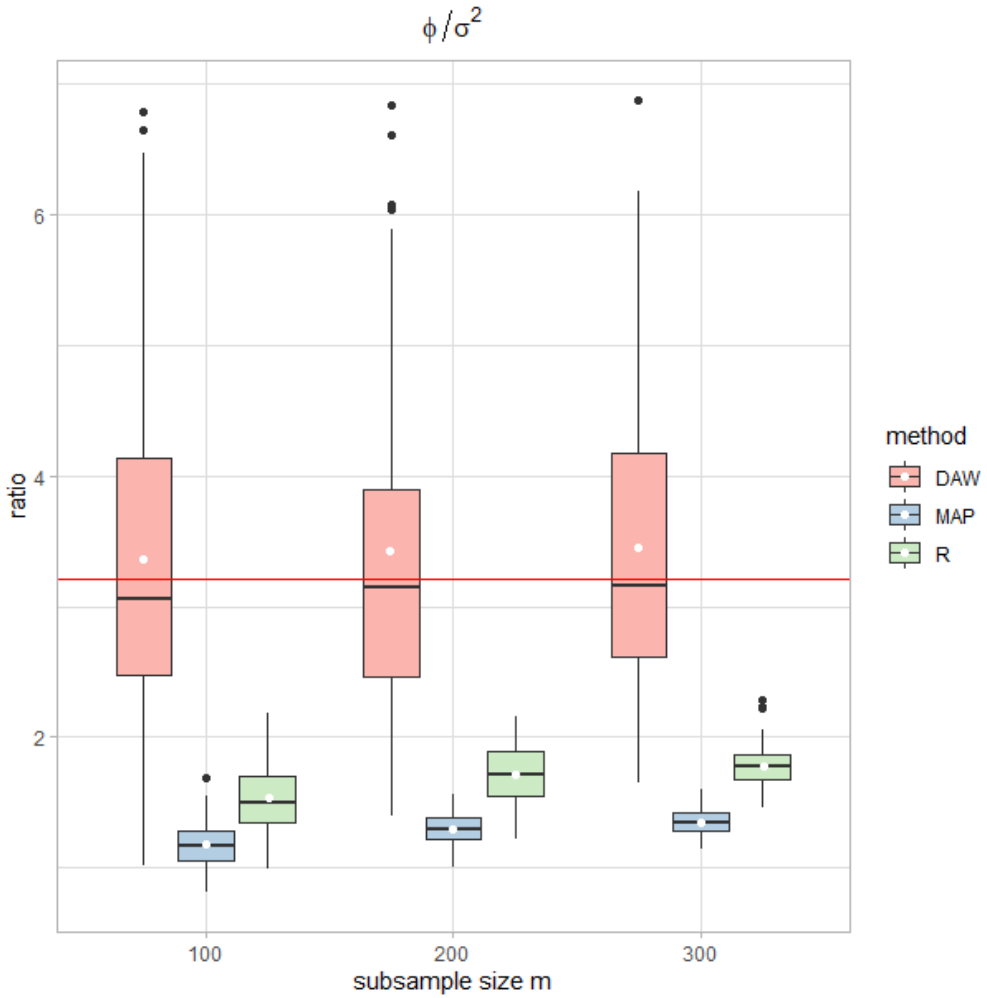

Figure: Boxplot of ϕ/σ^2 estimated by various subsampling methods under m=100,200 and 300.

Parameter Estimation: Nugget Effect

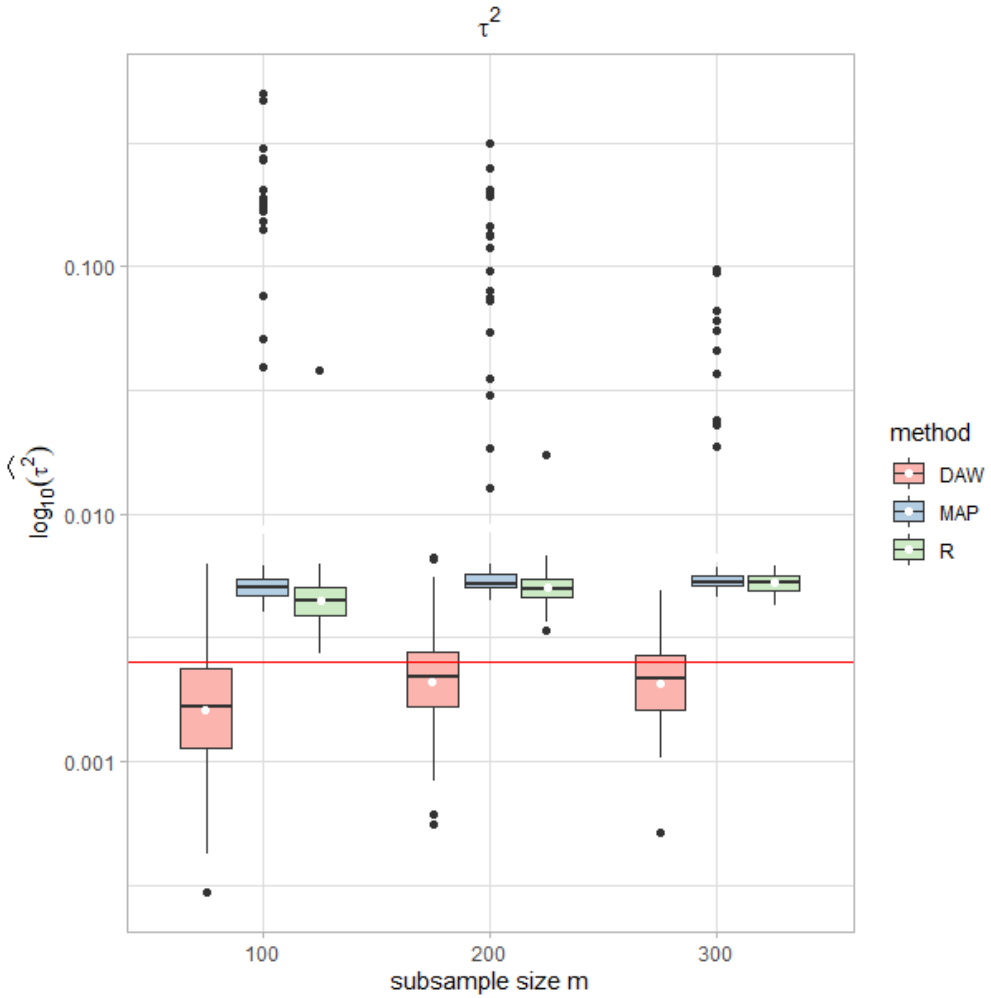

Figure: Boxplot of τ^2 estimated by various subsampling methods under m=100,200 and 300.

Parameter Estimation: Total MSEs

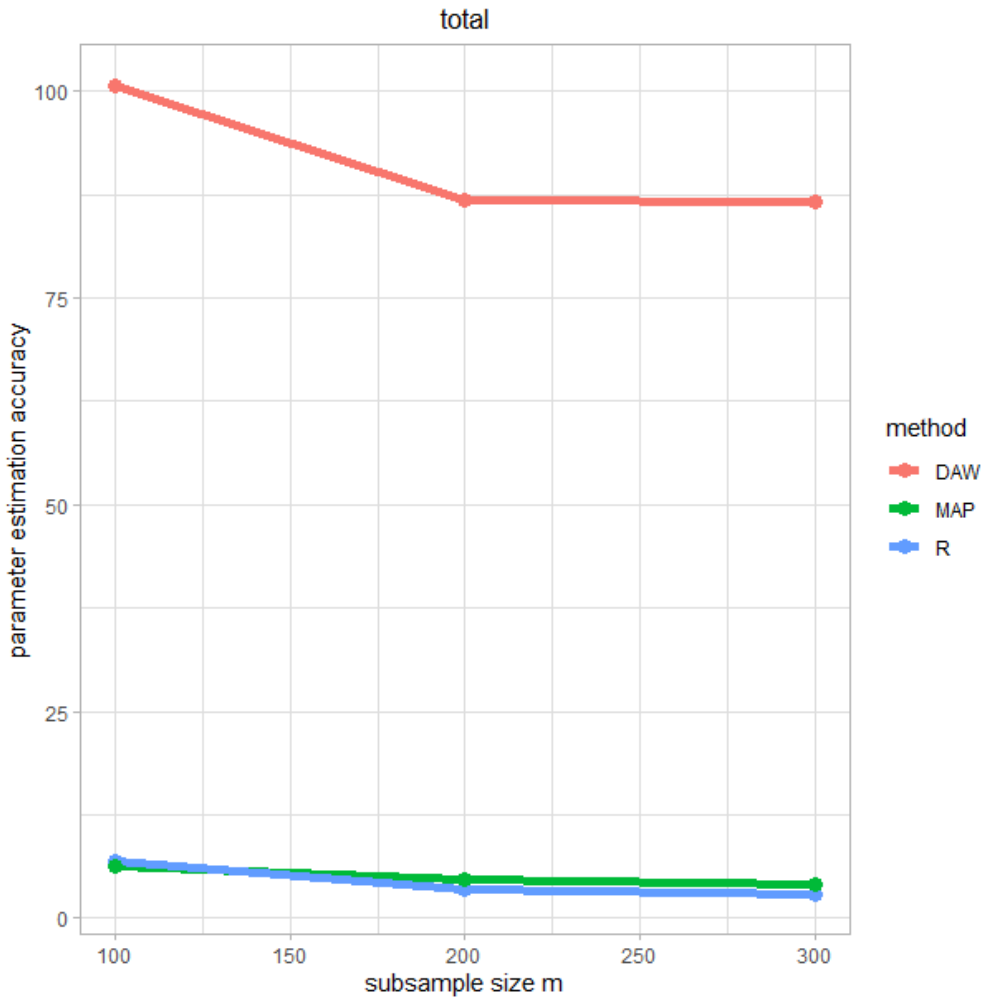

Figure: The sum of all parameters' MSEs

Prediction

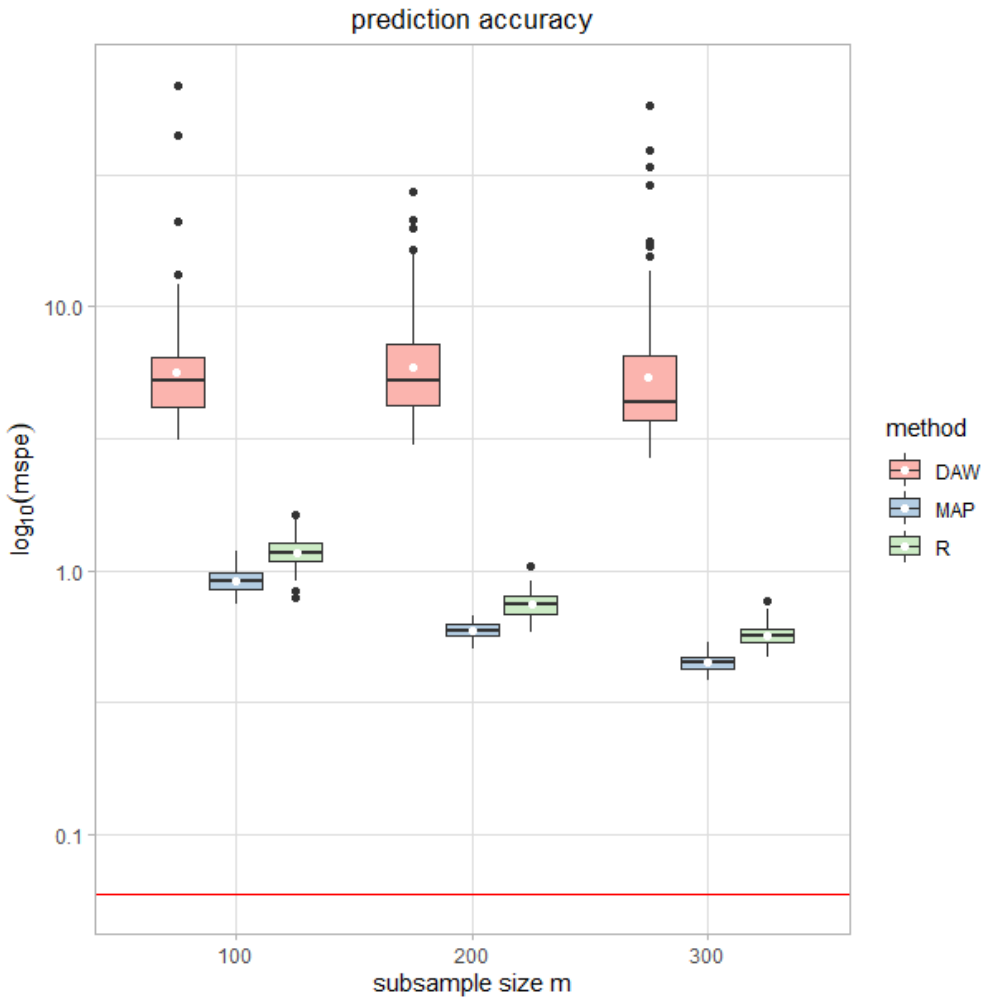

Figure: Boxplot of prediction accuracy of various subsamples.

Fill in Missing Values

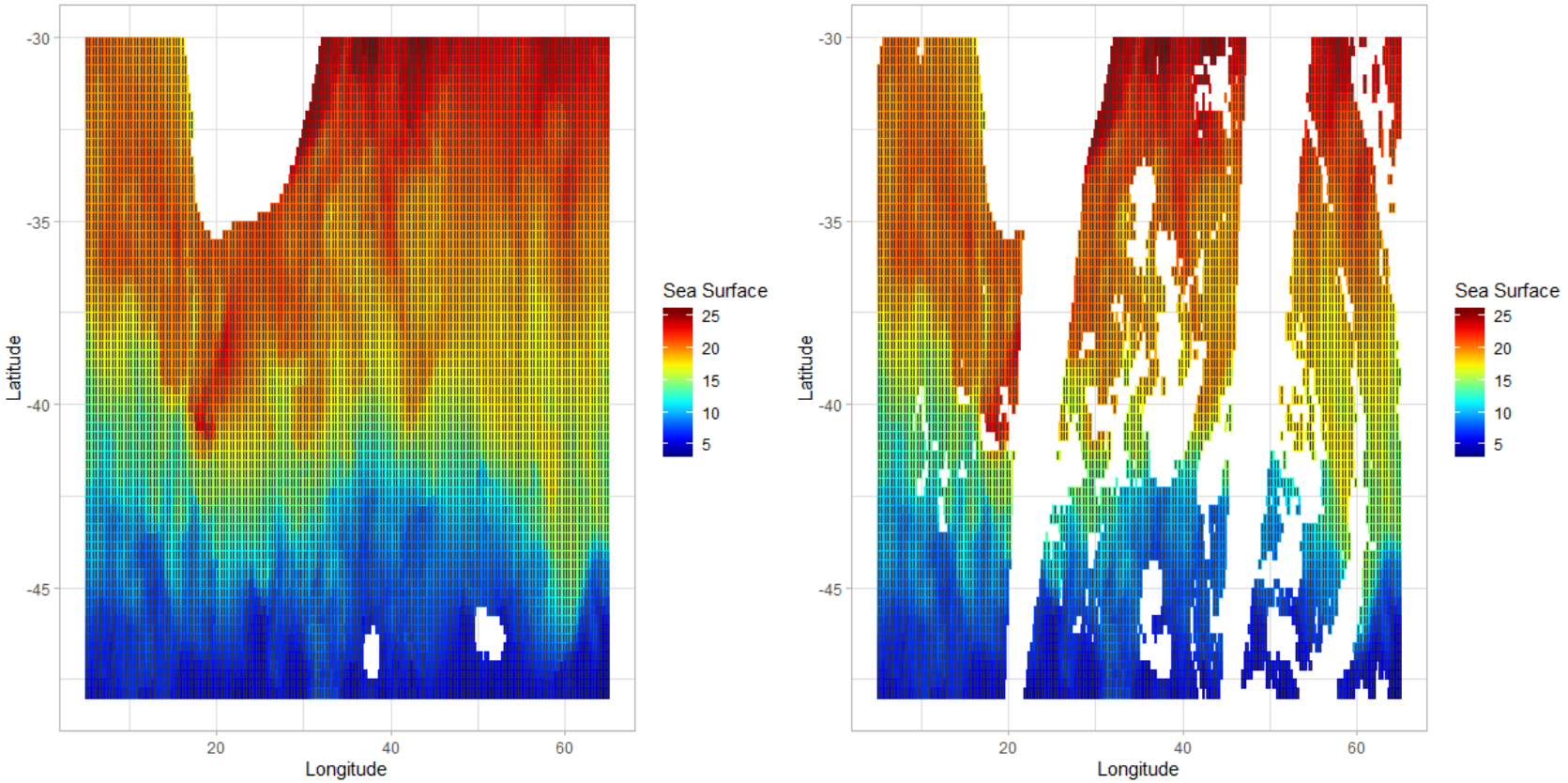

Figure: Sea surface temperatures on Day 10.

Subsampling Method

- MaxPro:
  - trend parameters, prediction.
  - computational limitation.
- Deep and Wide:
  - covariance parameters
  - large variance $\rightarrow$ ensemble
- Random:
  - is close to MaxPro.
  - can be used when we want to fill in missing values quickly and easily.
- Missing values in data of other days can be filled in by MaxPro.
- The selection of subsampling method based on data and our target.