Proposal

YanSong

Getting Started

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. In this project, we ignore the temporal component of the dataset and analyse temperatures on a Day 10. The data contains a lot of missing values which caused by land, satellite's orbital clipping and cloud cover. Our ultimate goal is to predict the missing values present in the Day 10 data with Gaussian geostatistical model.

However, Gaussian geostatistical model need $O(n^3)$ numerical operations to estimate parameters and make a prediction. It's time consuming. Thus, we consider a data reduction. We want to know if we can use only a small part of data to estimate the parameters and predict missing values, and how many subsamples we need to reach a satisfied results?

Data Information

Import Data

We use function readMat from the package R.matlab to convert the data from Matlab format to R.

```
library("R.matlab")
SST=readMat("F:/data/data_SST.mat",maxLength = NULL,fixNames = TRUE)
```

Data Discription

We can see that SST is a list with 4 variables. The lon.zone, lat.zone and time.period store longitude, latitude and time values, respectively. The longitude range from 5.125 to 64.875. The latitude range from -47.88 to -30.12. The main variable, SST.zone.period, is a three dimensional $72 \times 240 \times 331$ matrix (latitude, longitude, day) of sea surface temperatures given in degrees Celsius.

```
class(SST)
#[1] "list"
```

```
names(SST)
#[1] "lon.zone"        "lat.zone"        "time.period"
#[4] "SST.zone.period"
```

```
dim(SST$SST.zone.period)
#[1]  72 240 331
```

```
summary(SST$lon.zone)[c(1,4,6)]
#[1] "Min.   : 5.125  " "Mean   :35.000  " "Max.   :64.875  "
```

```
summary(SST$lat.zone)[c(1,4,6)]
#[1] "Min.   :-47.88  " "Mean   :-39.00  " "Max.   :-30.12  "
```

```
head(SST$time.period)
#       [,1]
#[1,] 733408
#[2,] 733409
#[3,] 733410
#[4,] 733411
#[5,] 733412
#[6,] 733413
```

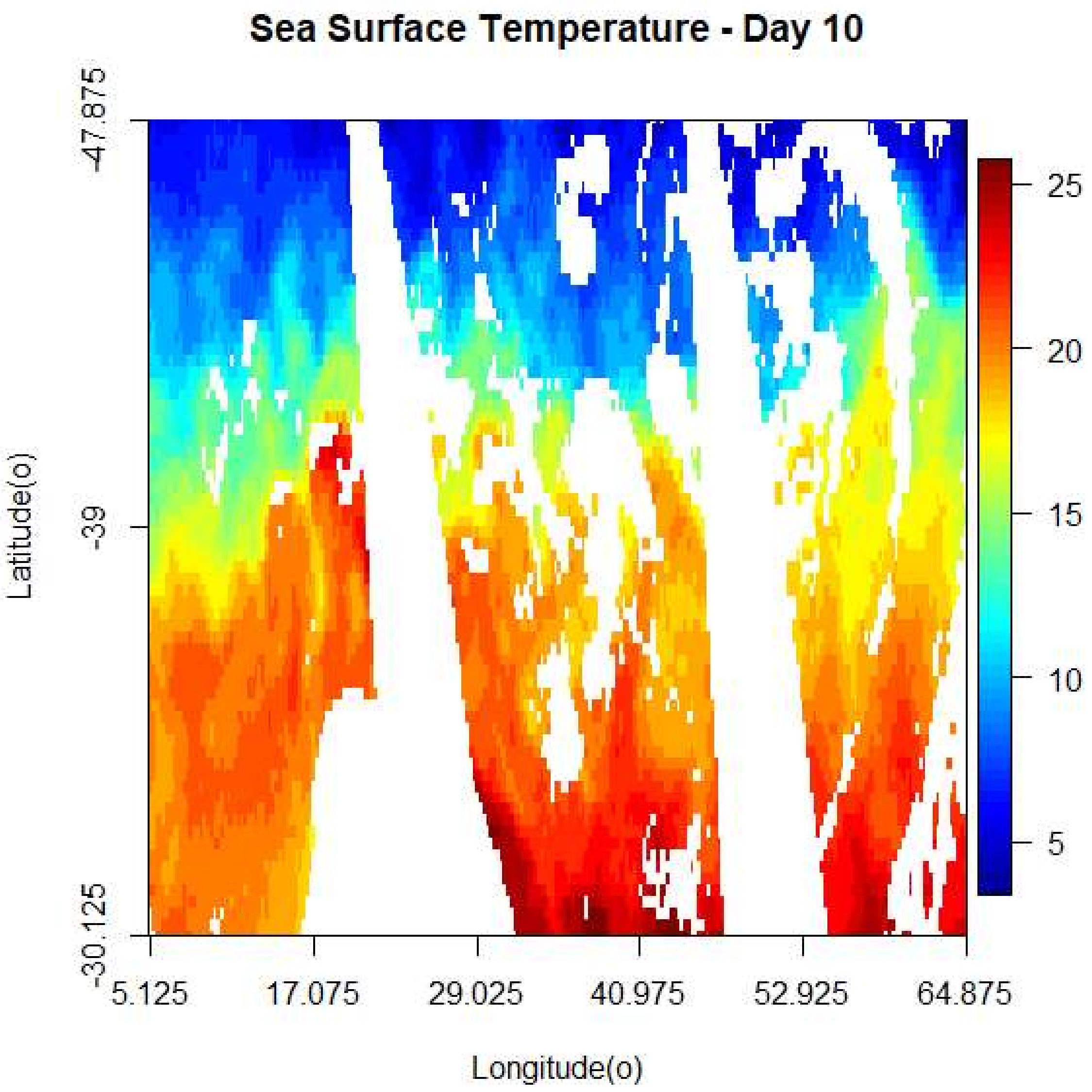
```
SST$SST.zone.period[1:5,1:5,10]
#          [,1]     [,2]     [,3]     [,4]     [,5]
#[1,]      NaN      NaN 20.33999 20.31999 20.39999
#[2,]      NaN 20.18999 20.20999 20.13999 20.32999
#[3,]      NaN 20.04999 20.16999 20.17999 20.18999
#[4,] 20.18999 20.09999 20.15999 20.24999 19.84999
#[5,] 20.41999 20.32999 20.07999 20.04999 19.74999
```

Exploratory Data Analysis

Data Visualization

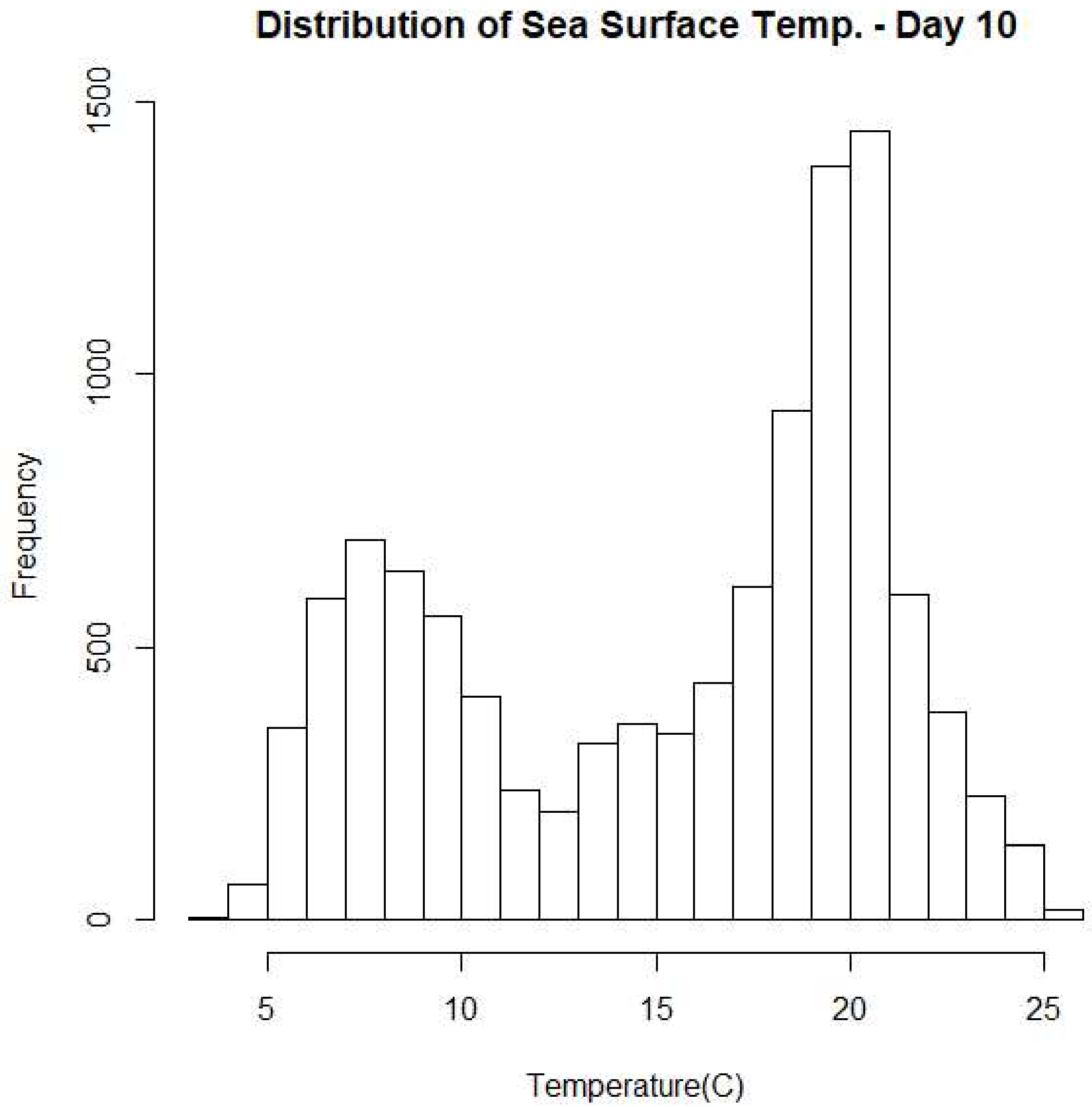
We show the sea surface temperatures recorded on day 10 by image.plot command in the fields package. We can see that there are numerous gaps in the data, corresponding to three main causes: land, satellite orbital clipping and cloud cover. A pronounced temperature gradient is visible from highs of over 25 degree in the north of the study area to a low of less 5 degree towards the southern boundary. It tells us that there is spatial correlation in the dataset and the data are not stationary because the mean temperature vary strongly with latitude.

```
lat=SST$lat.zone
lon=SST$lon.zone
par(mar=c(5,5,5,5))
image(t(SST$SST.zone.period[,,10]),col=tim.colors(25),xaxt="n",yaxt="n",ylim=c(0,1),
      main="Sea Surface Temperature - Day 10",xlab="Longitude(o)",ylab="Latitude(o)")
axis(1, at = seq(0, 1, by = 1/5), labels=seq(min(lon), max(lon),(max(lon)-min(lon))/5))
axis(2, at = seq(0, 1, by = 1/2), labels=rev(seq(min(lat), max(lat),(max(lat)-min(lat))/2)))
image.plot(SST$SST.zone.period[,,10], legend.only=T)
```



We then generate a histogram of sea surface temperatures for Day 10. The temperatures have a bimodal distribution, with a warm peak and a smaller cool peak. It may be caused by the large regions of warm and cool waters at the north and south of the study area.

```
hist(SST10,sqrt(240*2),main="Distribution of Sea Surface Temp. - Day 10",xlab = "Temperature (C)")
```



Model Selection

Consider a Gussian geostatistical model:

$$y(s_i) = \mu(s_i) + w(x_i) + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d}}{\sim} \mathrm{N}(0, \tau^2)$$

where $\mu(s_i)$ denotes the mean of $y(s_i)$, $\{w(s_i)\}$ denotes a spatial Gaussian process with $\mathrm{E}(w(s_i)) = 0$, $\mathrm{Var}(w(s_i)) = \sigma^2$ and $\mathrm{Corr}(w(s_i), w(s_j)) = \rho(\|s_i - s_j\|; \phi)$. The correlation function is chosen from a certain parametric family, such as the Matern, exponential, or spherical covariance models. τ^2 denotes the nugget variance. Under this model, $\{y(s)\}$ follows a multivariate Gaussian distribution as follows:

$$[y(s_1), \ldots, y(s_n)]^T \sim \mathrm{N}(\mu, \Sigma)$$

where $\mu = (\mu(s_1), \ldots, \mu(s_n))^T$, $\Sigma = \sigma^2 R + \tau^2 I$, I is the $n \times n$ identity matrix, and R is an $n \times n$ correlation matrix with the (i, j)th element $\rho(\|s_i - s_j, \phi\|)$.

The semivariogram is given by $\gamma(h) = \sigma^2(\rho(0) - \rho(h))$. For Matren class, the semivariogram is:

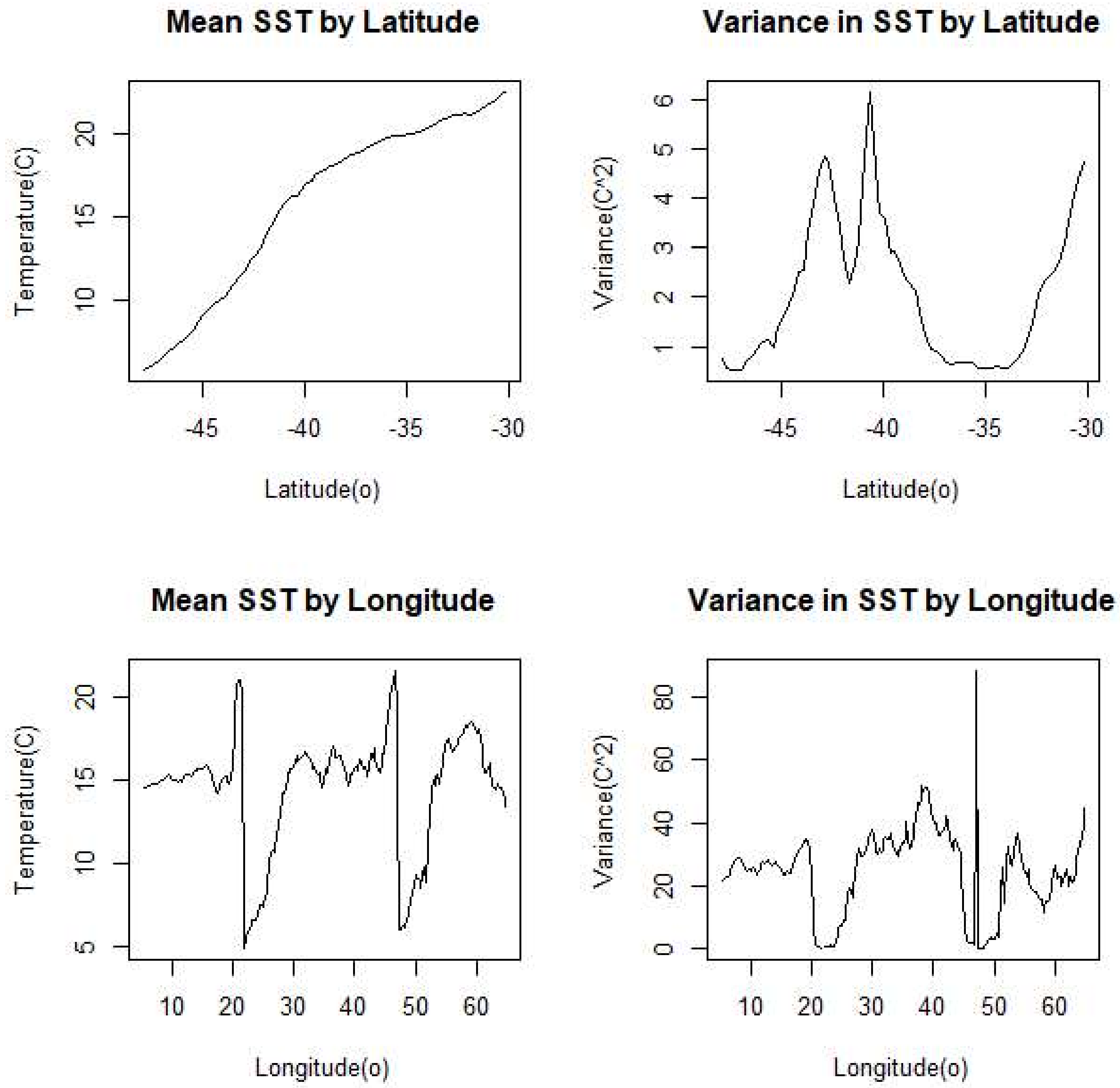
$$\gamma(h; \theta) = \sigma^2 \big(1 - \frac{(h/\theta)^\nu \mathcal{K}_\nu(h/\theta)}{2^{\nu-1}\Gamma(\nu)}\big)$$

where $\mathcal{K}_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν.

Trend Selection

The non-stationarity of the sea surface temperatures is also shown in following figures. As latitudes approach the equator, the mean temperature rises. Variances are highest in the zone of mixing between -44N and -38N. However, both mean and variance are relatively constant across longitudes, with large deviations on both plots caused by the missing values.

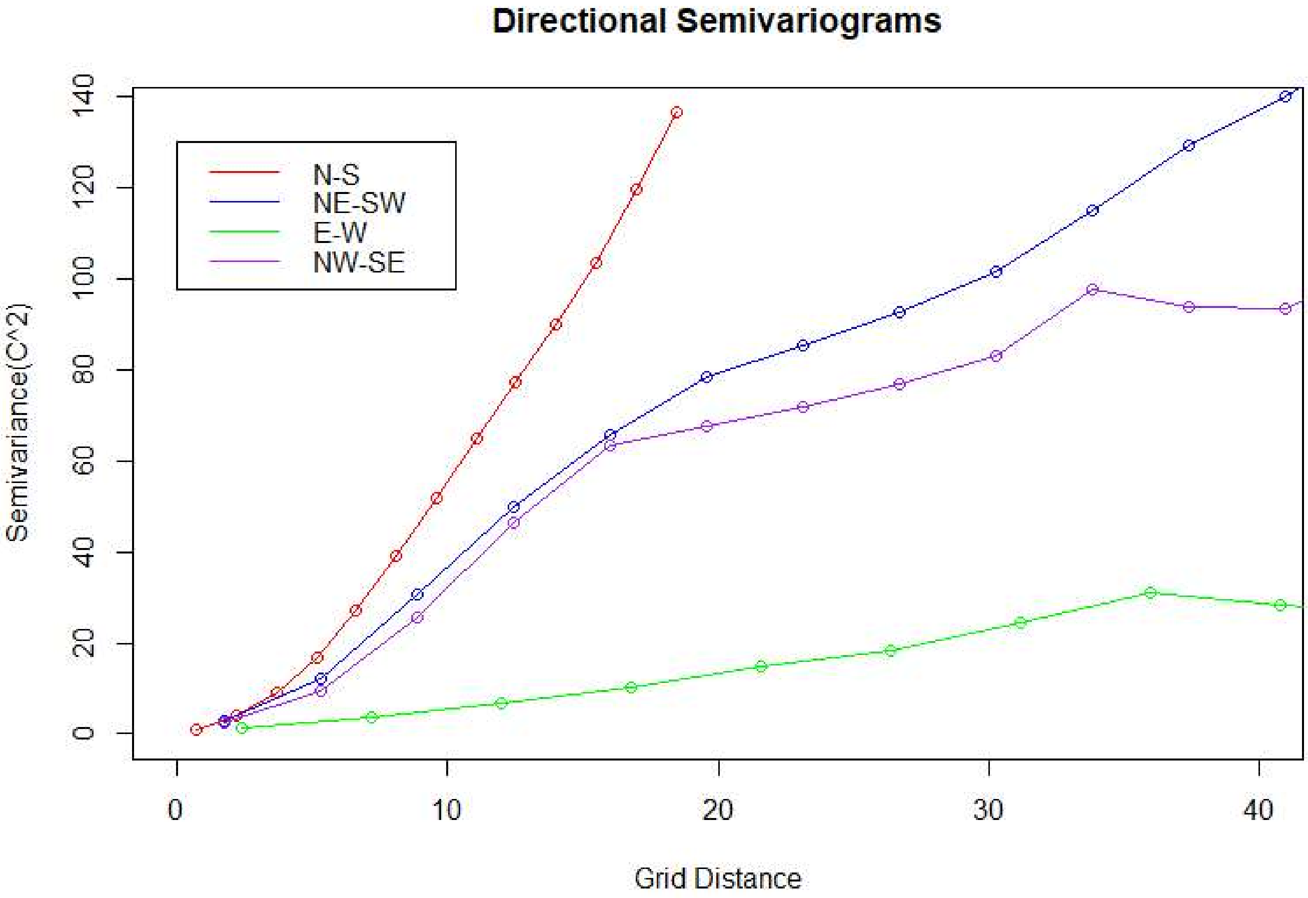
```
par(mfrow=c(2,2))
#### mean & latitude
latMean=rowMeans(SST10,na.rm = TRUE)
plot(lat,latMean,type="n",main="Mean SST by Latitude",ylab="Temperature(C)",xlab = "Latitude(o)")
lines(lat,latMean)
#### variance & latitude
var1=function(x){return(var(x,na.rm = TRUE))}
latVar=apply(SST10,1,var1)
plot(lat,latVar,type="n",main="Variance in SST by Latitude",ylab="Variance(C^2)",xlab = "Latitude(o)")
lines(lat,latVar)
#### mean & longitude
lonMean=colMeans(SST10,na.rm = TRUE)
plot(lon,lonMean,type="n",main="Mean SST by Longitude",ylab="Temperature(C)",xlab = "Longitude(o)")
lines(lon,lonMean)
#### var & longitude
lonVar=apply(SST10,2,var1)
plot(lon,lonVar,type="n",main="Variance in SST by Longitude",ylab="Variance(C^2)",xlab = "Longitude(o)")
lines(lon,lonVar)
```



We can confirm that the data are non-stationary by plotting directional variograms. As can be seen in the figure, sea surface temperatures are distinctly non-stationary in the N-S direction. The semivariogram for these relationships approximates a linear trend and does not reach a sill. This non-staionary also influence NW-SE and NE-SW variograms. Only the E-W variances form something like a stationary variogram, reaching a rough sill of around 30 degree.

```
library("geoR")
loc.x1=rep(lon,each=72)
loc.x2=rep(lat,times=240)
y=c(SST10)
geoSST10=as.geodata(cbind(loc.x1,loc.x2,y))

Variog=variog4(geoSST10,direction = c(0,pi/4,pi/2,3*pi/4))
plot(Variog$`0`,main="Directional Semivariograms",xlab="Grid Distance",ylab="Semivariance(C^
2)",xlim=c(0,40))
lines(Variog$`0`,col="red")
lines(Variog$`45`,col="blue")
lines(Variog$`90`,col="green")
lines(Variog$`135`,col="purple")
legend(0,130,c("N-S","NE-SW","E-W","NW-SE"),lty=c(1,1,1,1),lwd=c(1,1,1,1),col=c("red","blue","g
reen","purple"))
```



As we saw in above figures, there is a clear trend between mean temperature and latitude. So we will use a linear trend.

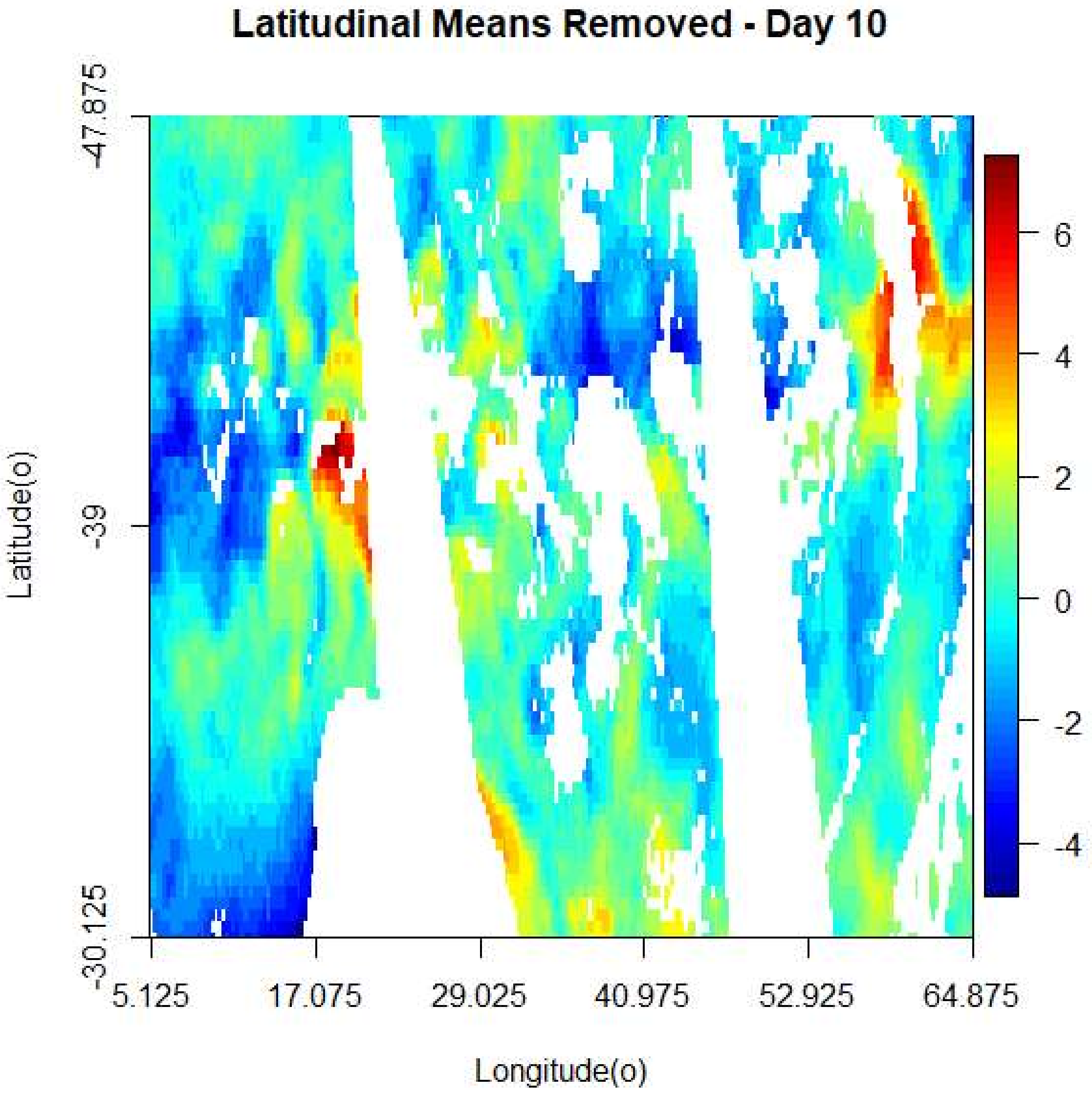
Variance Function Selection

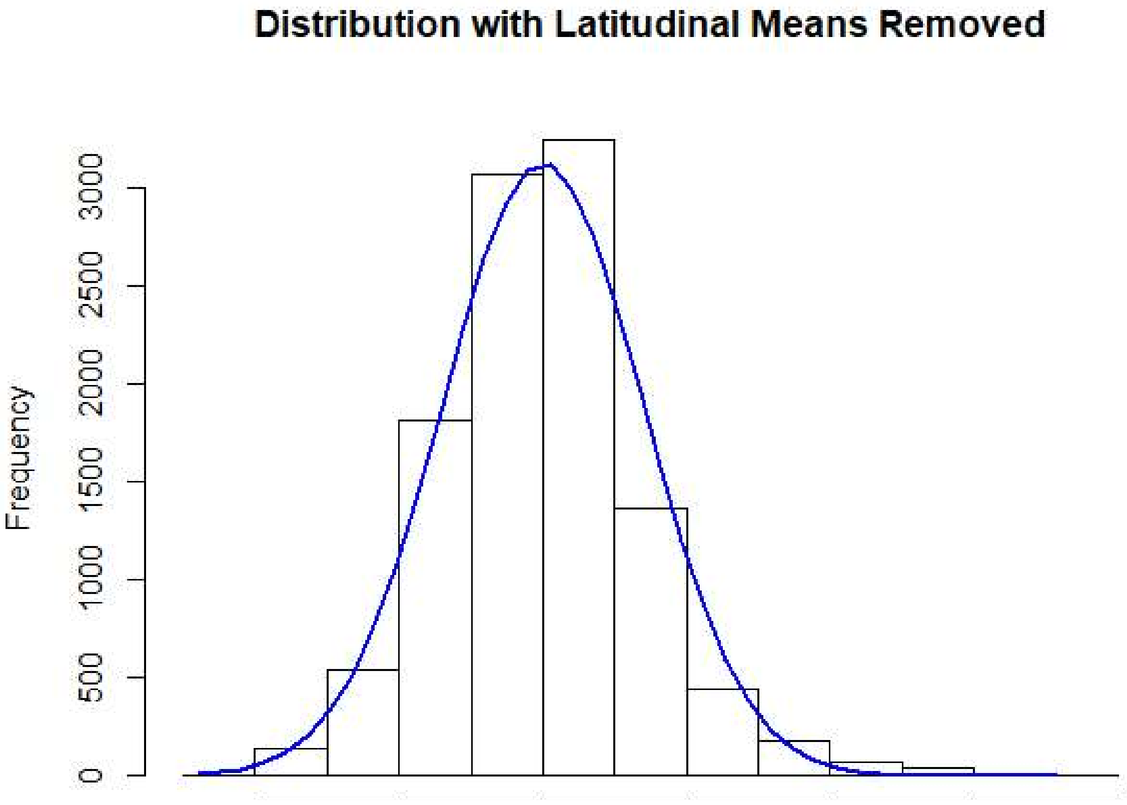
Next, we will choose a variance function. Instead of modeling this trend, since we have a mean temperature value for each latitude, we can simply subtract these latitudinal means from the Day 10 data before generating the semivariogram.

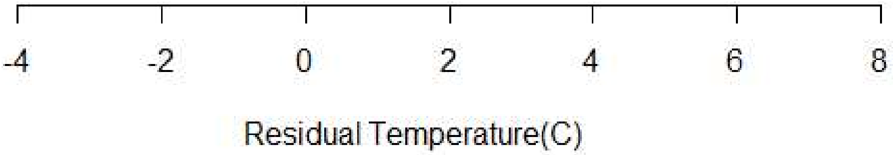
We can see from the plots of the residuals that there are still regions of high and low temperatures and evidence of structure. But the residuals have a nearly Gaussian distribution.

```
latmean=rep(latMean,times=240)
y_latm=y-latmean
SST10latm=cbind(loc.x1,loc.x2,y_latm)
SST10_latm=matrix(y_latm,nrow=72,ncol=240)
par(mar=c(5,5,5,5))
image(t(SST10_latm),col=tim.colors(25),xaxt="n",yaxt="n",ylim=c(0,1),
      main="Latitudinal Means Removed - Day 10",xlab="Longitude(o)",ylab="Latitude(o)")
axis(1, at = seq(0, 1, by = 1/5), labels=seq(min(lon), max(lon),(max(lon)-min(lon))/5))
axis(2, at = seq(0, 1, by = 1/2), labels=rev(seq(min(lat), max(lat),(max(lat)-min(lat))/2)))
image.plot(SST10_latm, legend.only=T)

h=hist(y_latm,main = "Distribution with Latitudinal Means Removed",xlab="Residual Temperature
(C)")
xfit=seq(min(y_latm,na.rm = TRUE),max(y_latm,na.rm = TRUE),length=40)
yfit=dnorm(xfit,mean=mean(y_latm,na.rm = TRUE),sd=sd(y_latm,na.rm = TRUE))
yfit=yfit*diff(h$mids[1:2])*(length(y_latm)-6335)
lines(xfit,yfit,col="blue",lwd=2)
```

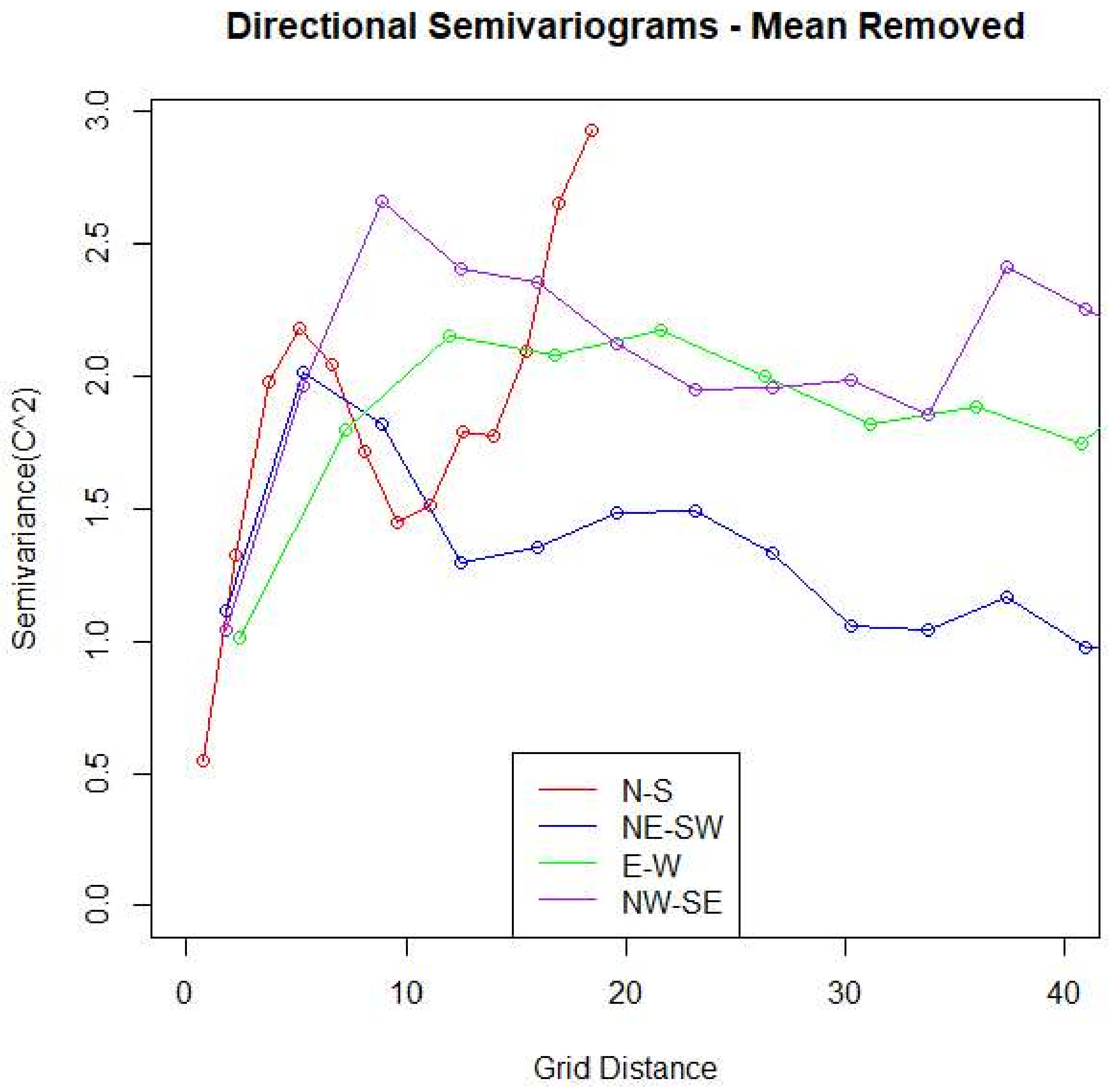

Latitudinal Means Removed - Day 10
Latitude(o)
-47.875
-39
-30.125
Longitude(o)
5.125
17.075
29.025
40.975
52.925
64.875
6
4
2
0
-2
-4


Distribution with Latitudinal Means Removed
Frequency
0
500
1000
1500
2000
2500
3000



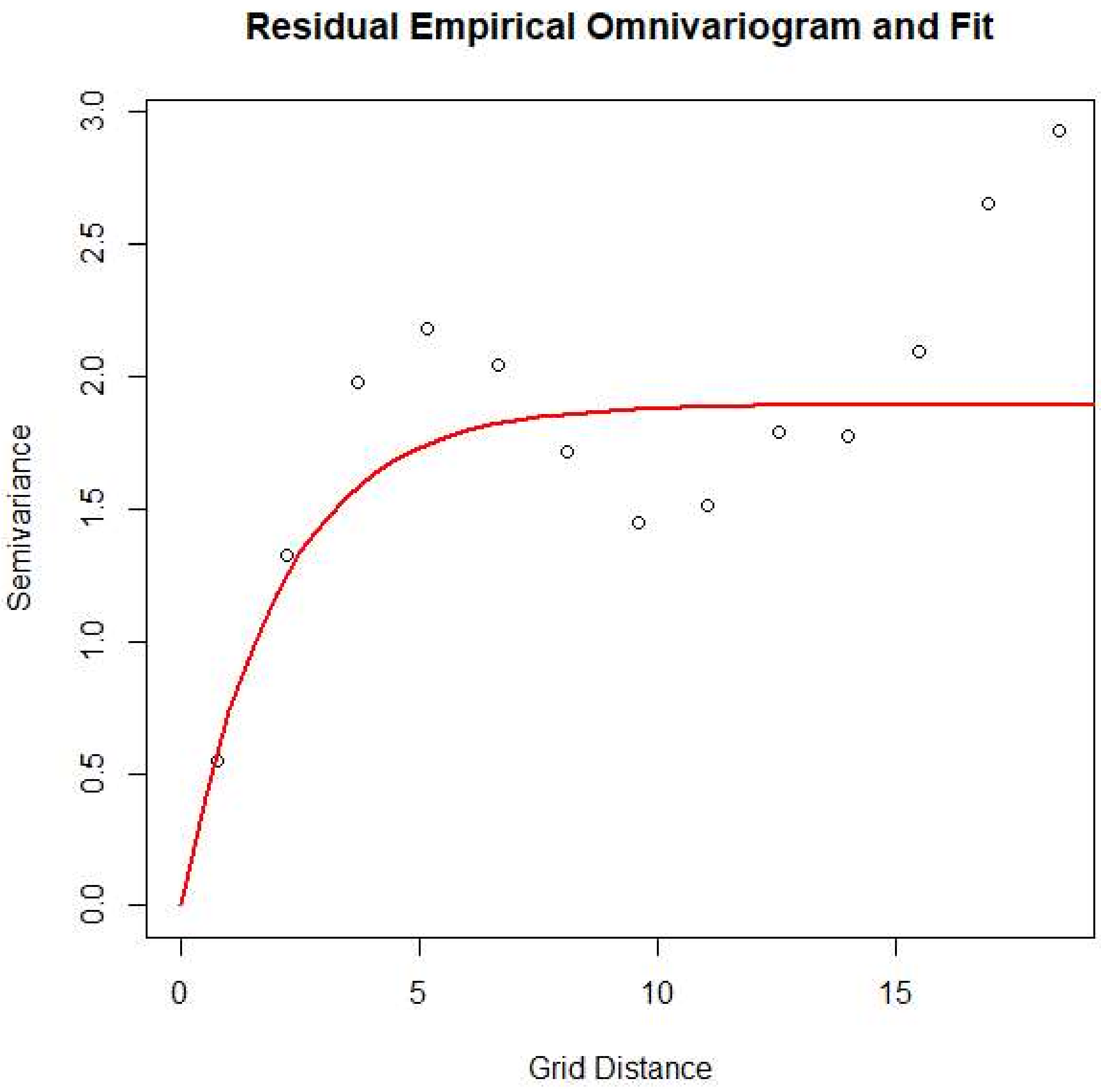
The empirical directional variograms for the residual data can be seen in following figure. They are much closer to stationary now. The N-S direction still has a smaller range than E-W which means a weaker correlation in this direction.

```
geoNew=as.geodata(SST10latm)
VariogNew=variog4(geoNew,direction = c(0,pi/4,pi/2,3*pi/4))
plot(VariogNew$`0`,main="Directional Semivariograms - Mean Removed",xlab="Grid Distance",ylab
="Semivariance(C^2)",xlim=c(0,40))
lines(VariogNew$`0`,col="red")
lines(VariogNew$`45`,col="blue")
lines(VariogNew$`90`,col="green")
lines(VariogNew$`135`,col="purple")
legend("bottom",c("N-S","NE-SW","E-W","NW-SE"),lty=c(1,1,1,1),lwd=c(1,1,1,1),col=c("red","blu
e","green","purple"))
```



Finally, we fit the variogram with exponential variance function. It is a special case of matern class and have a common use. We can see that it has a good performance in our data so we will use it in our following simulation.

```
Var0=VariogNew$`0`
geoNewS=sample.geodata(geoNew,500)
VarS=variog(geoNewS)
fitVAR=variofit(VarS,ini=c(0.5,1),cov.model = "exponential",fix.nugget = FALSE,nugget=1,max.dist = 50)
plot(Var0,main = "Residual Empirical Omnivariogram and Fit",ylab="Semivariance",xlab="Grid Distance")
lines(fitVAR,col="red",lwd=2)
```



Future Work

1. We select m subsamples with following three different subsampling schemes under different m:

1. Random: m subsamples are chosen randomly from original data.
2. Deep and Wide: 5 center points are selected together with their nearest $\frac{m}{5}$ samples.
3. MaxPro: a kind of design method with space-filling property.

We fit these three subsamples to Gaussian model we selected. For each subsample, we can obtain a set of parameter estimations. We plug the estimated parameters into Kriging model to predict the missing values. We replicate the above process 100 times.

2. We use results from “full data” to be true value and calculate MSE of three kinds of subsamples under different m.
3. We analyse our results and answer our question.