2016 NBA Finals

Saba Paya

March 23, 2018

Abstract:

Geostatistics is the estimation and prediction for spatially continuous phenomena using spatial locations. The phenomena is the distribution of regionalized varibales and is applied to numerous areas such as mining, epidemiology, ecology, and environmental science to name a few. In this project, I will be looking at NBA data and using the (x,y) coordinates of the basketball court.

In order to understand how the NBA can use spatial location to improve player performance, I will look at a target variable, whether a shot is made or missed (result), and two colocated variables, the distance the shot is taken from in feet (distance) and the type of shot taken (type). The purpose of this project is to look at spatial location and measure which kriging method is the best in making predictions of whether a shot is made or missed.

The data was obtained from the BigDataBall, a website that captures NBA play by play stats for every game. The sample data that I have used is a play by play from a 2016 Playoff Game between the Cleveland Cavaliers and Golden State Warriors.

In the first part, the data is cleaned and reduced from 44 variables to just three variables and the (x,y) coordinates. The variables with duplicate coordinates have their data averaged so that each coordiante has a unique value. Then, non-spatial analysis is done on each variable to check for normality before proceeding with the predictions.

Next, variogram fitting is done on the variables to see which model fits the data best. Once a decent model is fit on the data predictions are performed to determnine which kriging model is the best. The following predictions: ordinary, universal, cokriging, and indicator kriging are performed to help determine which works best. Ordinary kriging is when the mean is assumed to be constant and unknown. Universal kriging is when the mean is not constant and unknown. Cokriging is when there are colocated variables in addition to the target response. In this case, the ordinary and indicator predictions are the same as indicator kriging requires the the data to be converted to binary values of 0 and 1. Since, my target variable is already in the binary 0/1 format, there is no need to change the data and therefore the results are the same as ordinary kriging.

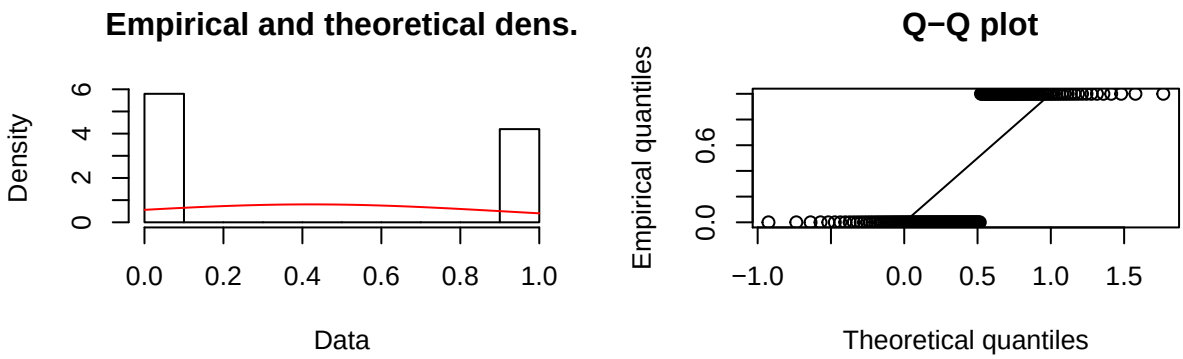
To determine which predictor performs best, cross validation is applied to each method and the one with the smallest predicted residual sum of squares is considered to be the best predictor. Through the analysis of this data, the best predictor is indicator kriging.

The Data

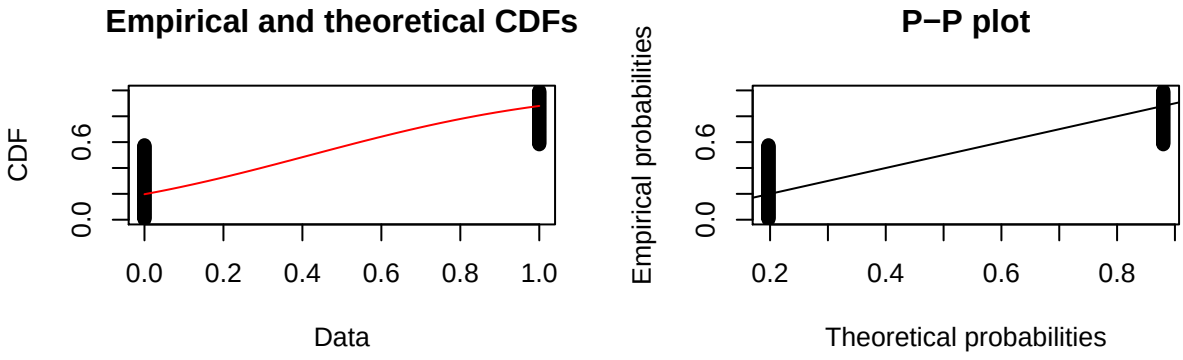
```
##      x    y result distance type
## 1 24.8  5.2      0        0    3
## 2 25.1 88.7      0        0    3
## 3 24.9 89.1      0        0    3
## 4 25.0  5.1      1        0   22
## 5 25.4  5.2      1        0   12
## 6 25.2 88.7      1        0   18
```

Non-Spatial Analysis

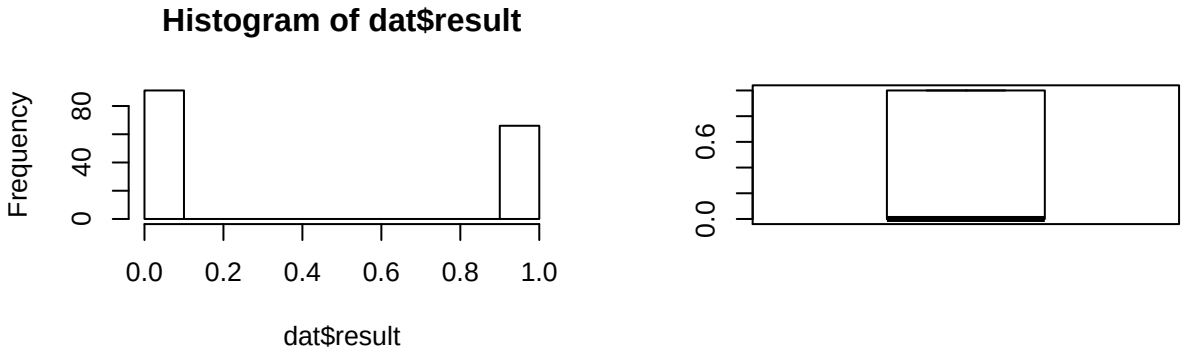
Result Variable — Binary Variable


Empirical and theoretical dens.
Density
0
2
4
6
0.0
0.2
0.4
0.6
0.8
1.0
Data

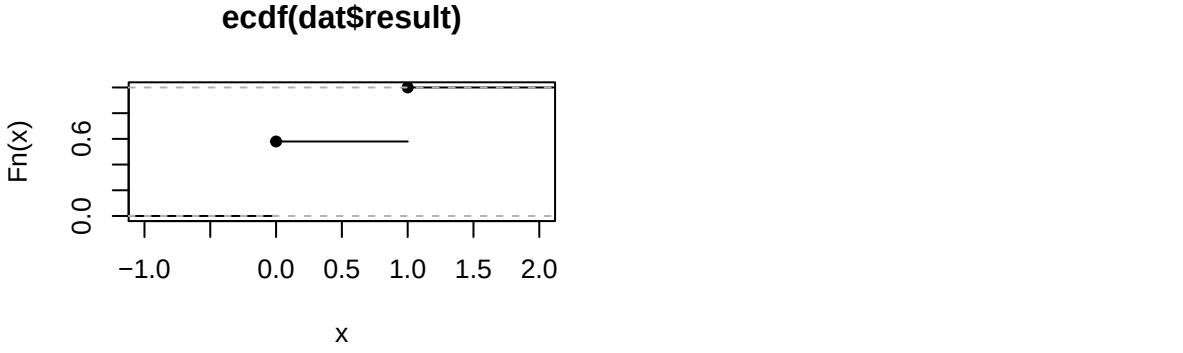
Q−Q plot
Empirical quantiles
0.0
0.6
−1.0
0.0
0.5
1.0
1.5
Theoretical quantiles


Empirical and theoretical CDFs
CDF
0.0
0.6
0.0
0.2
0.4
0.6
0.8
1.0
Data

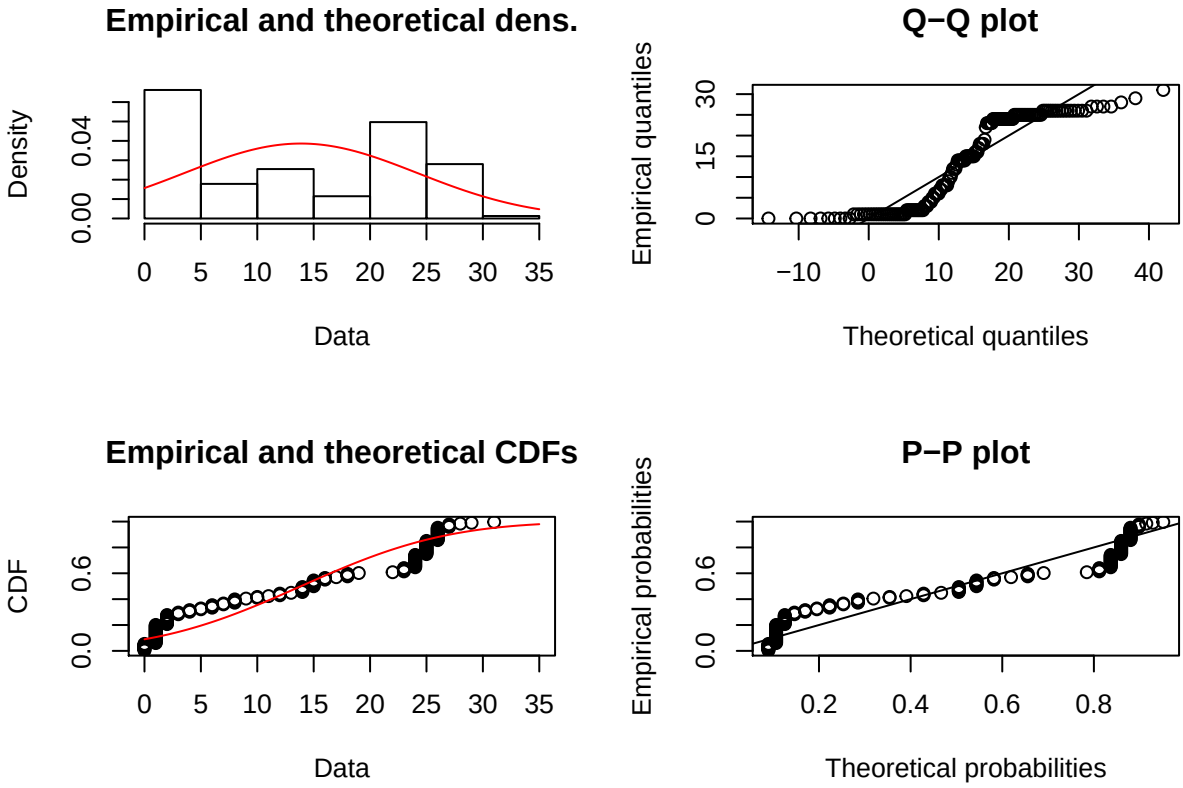
P−P plot
Empirical probabilities
0.0
0.6
0.2
0.4
0.6
0.8
Theoretical probabilities


Histogram of dat$result
Frequency
0
40
80
0.0
0.2
0.4
0.6
0.8
1.0
dat$result

0.0
0.6

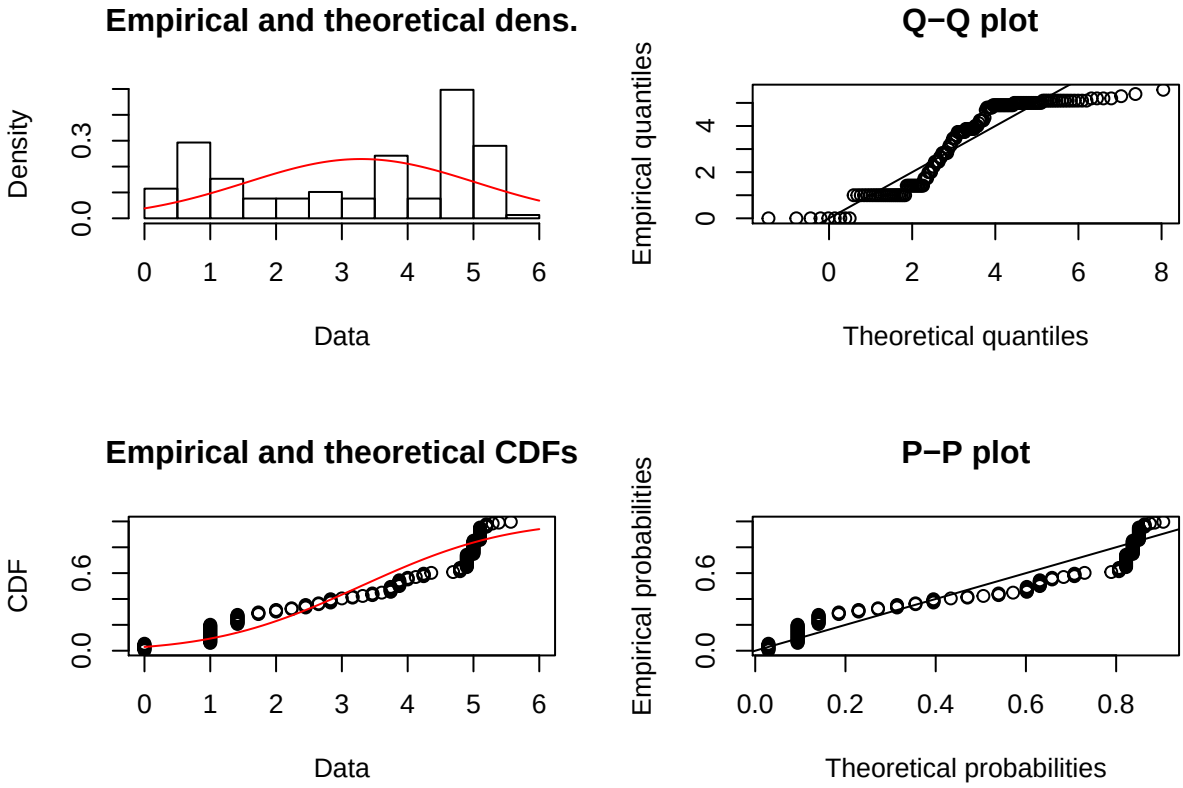

ecdf(dat$result)
Fn(x)
0.0
0.6
−1.0
0.0
0.5
1.0
1.5
2.0
x

Distance Variable

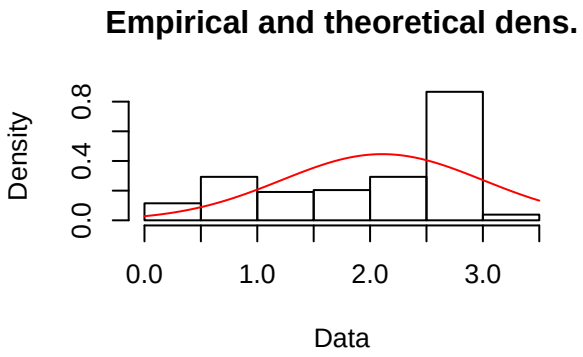


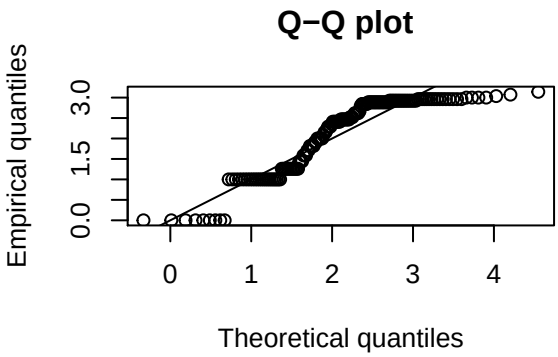
Performing transformations to see if they give more normal distribution

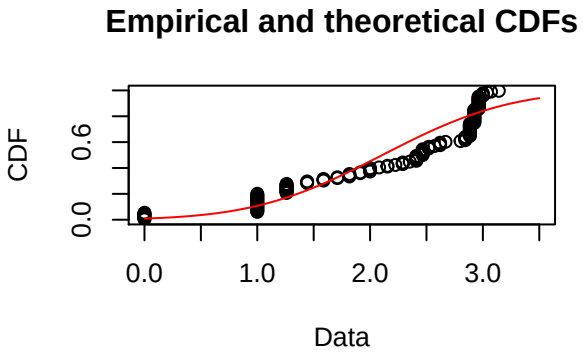
Applying a square root transformation

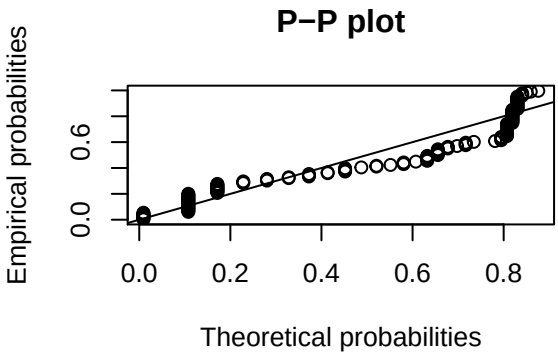


Transforming to the 1/3 power

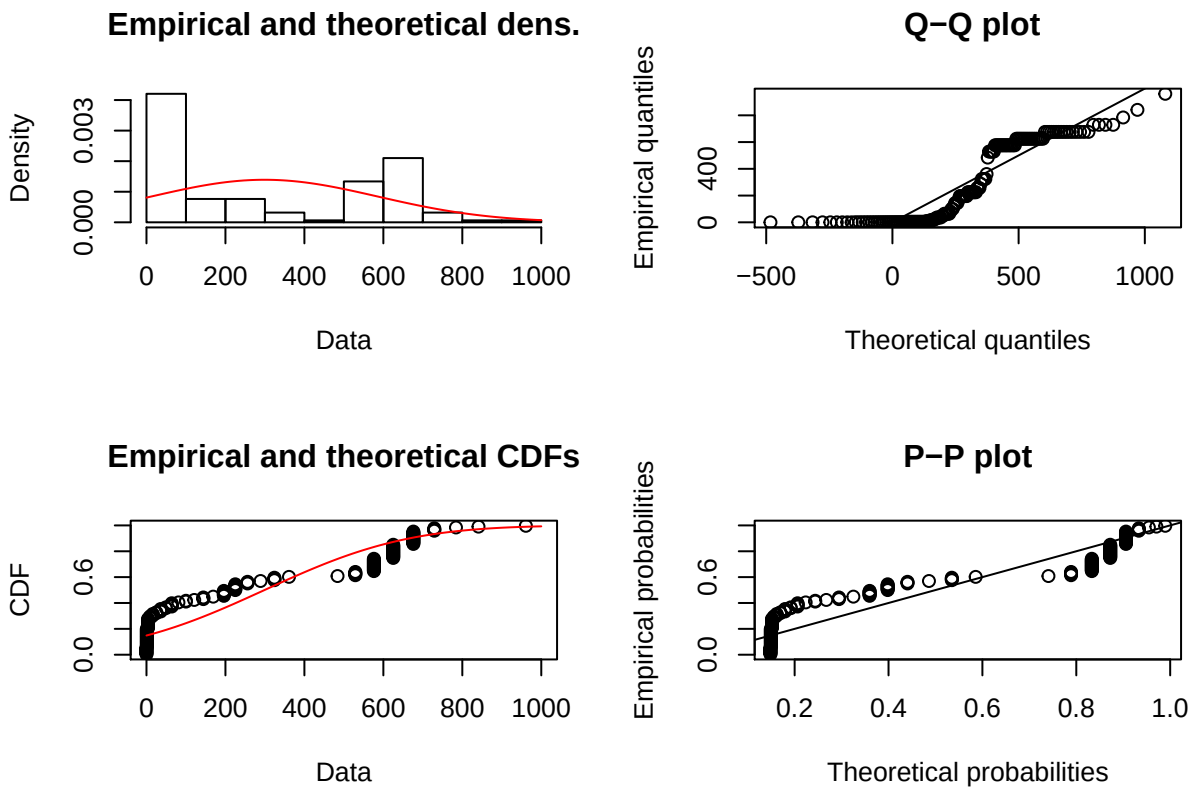






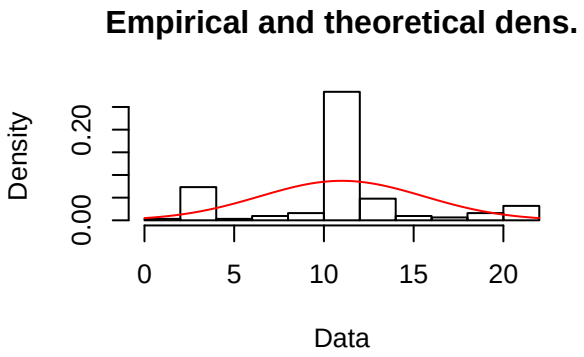


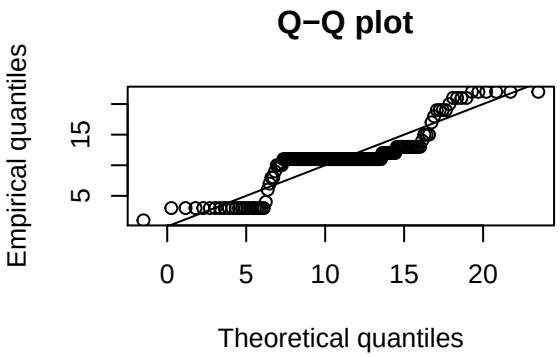
Appyling a square root transformation

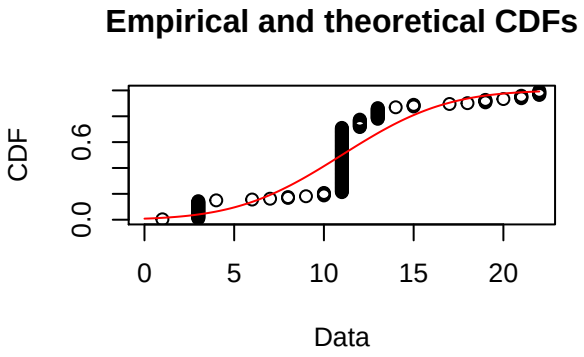


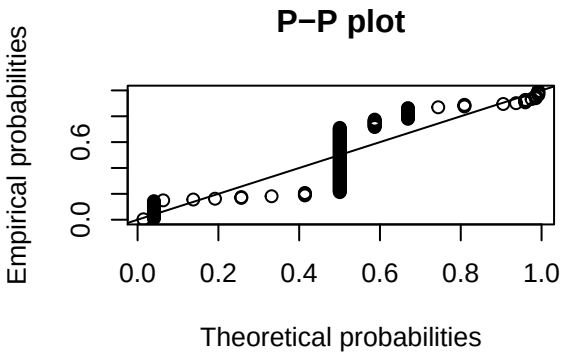
Squaring the distance variable gives the most normal distribution.

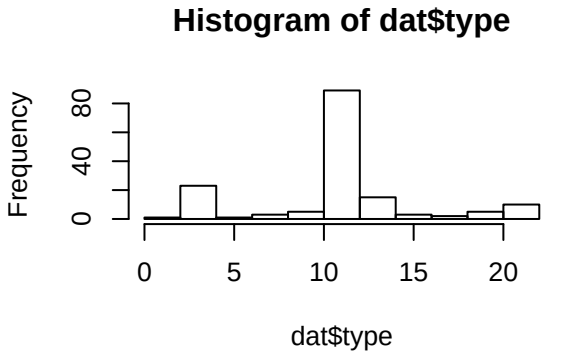
Type Variable

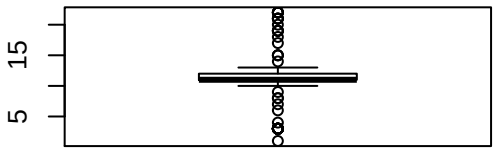

Empirical and theoretical dens.
Density
0.00
0.20
0
5
10
15
20
Data

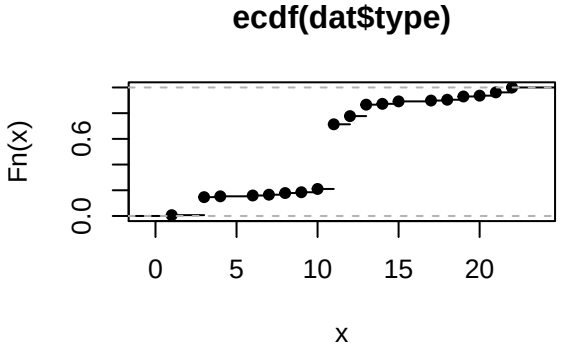

Q−Q plot
Empirical quantiles
5
15
0
5
10
15
20
Theoretical quantiles


Empirical and theoretical CDFs
CDF
0.0
0.6
0
5
10
15
20
Data


P−P plot
Empirical probabilities
0.0
0.6
0.0
0.2
0.4
0.6
0.8
1.0
Theoretical probabilities

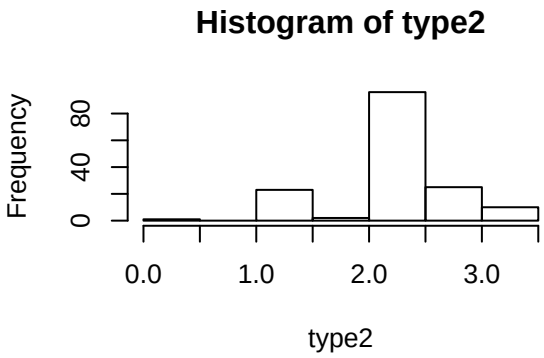

Histogram of dat$type
Frequency
0
40
80
0
5
10
15
20
dat$type

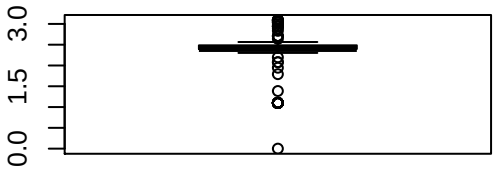

5
15

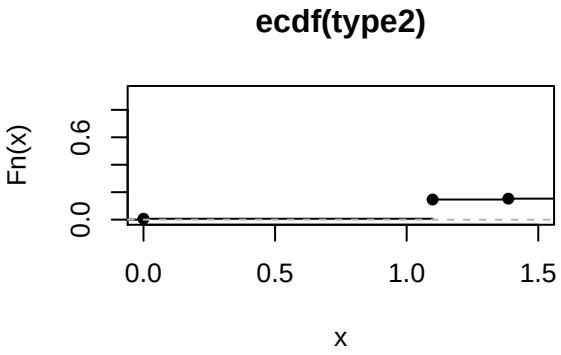

ecdf(dat$type)
Fn(x)
0.0
0.6
0
5
10
15
20
x

Performing transformations to give more normal distribution

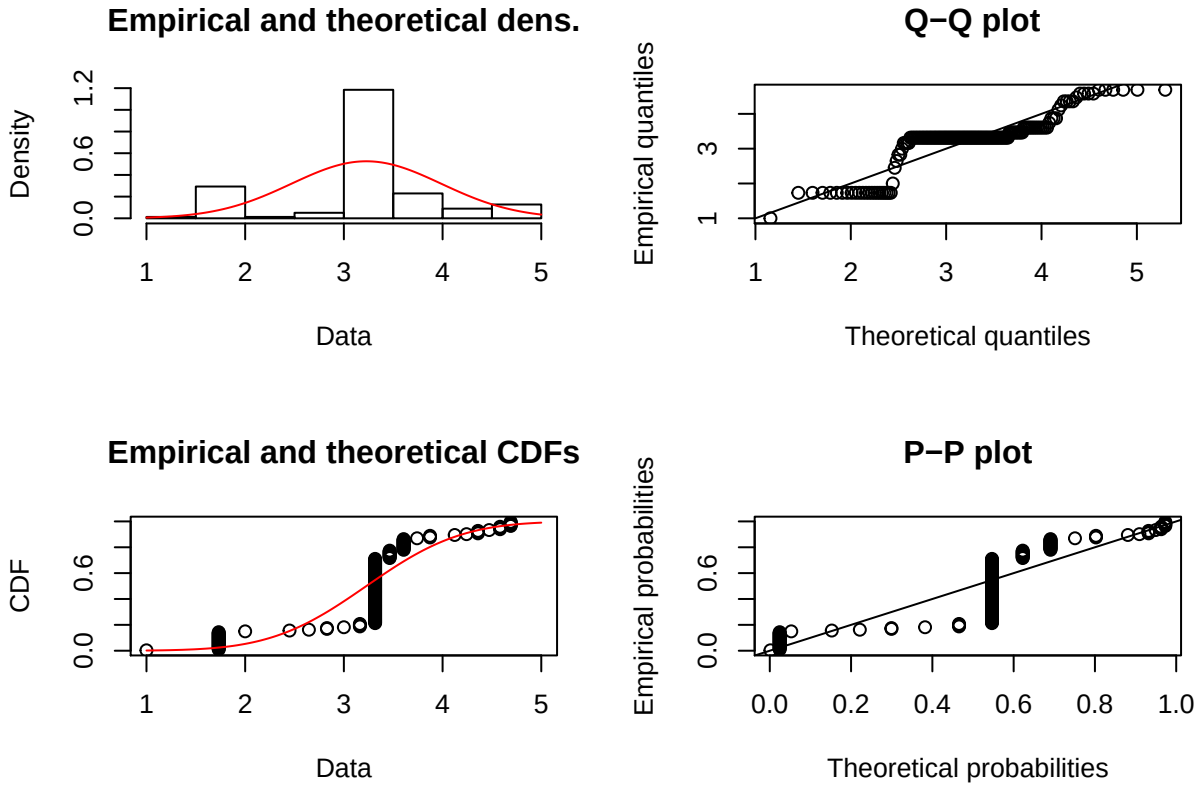
Applying a log transformation







Appyling a square root transformation

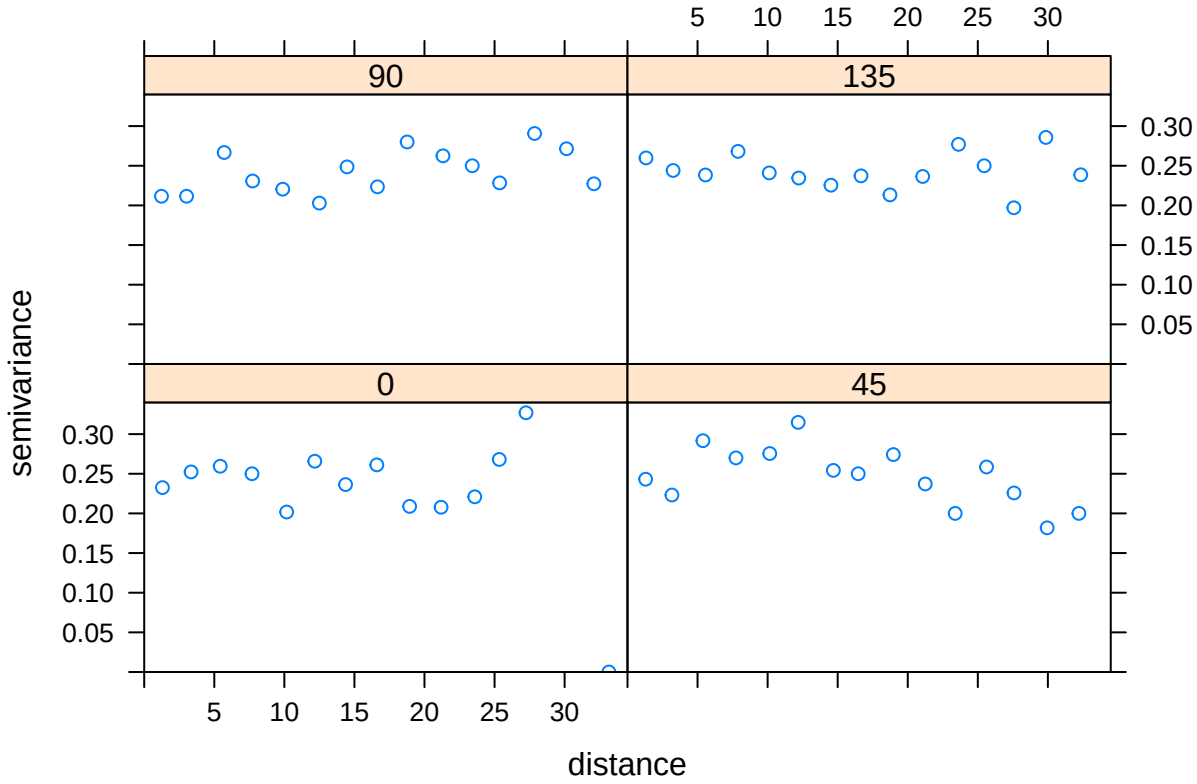


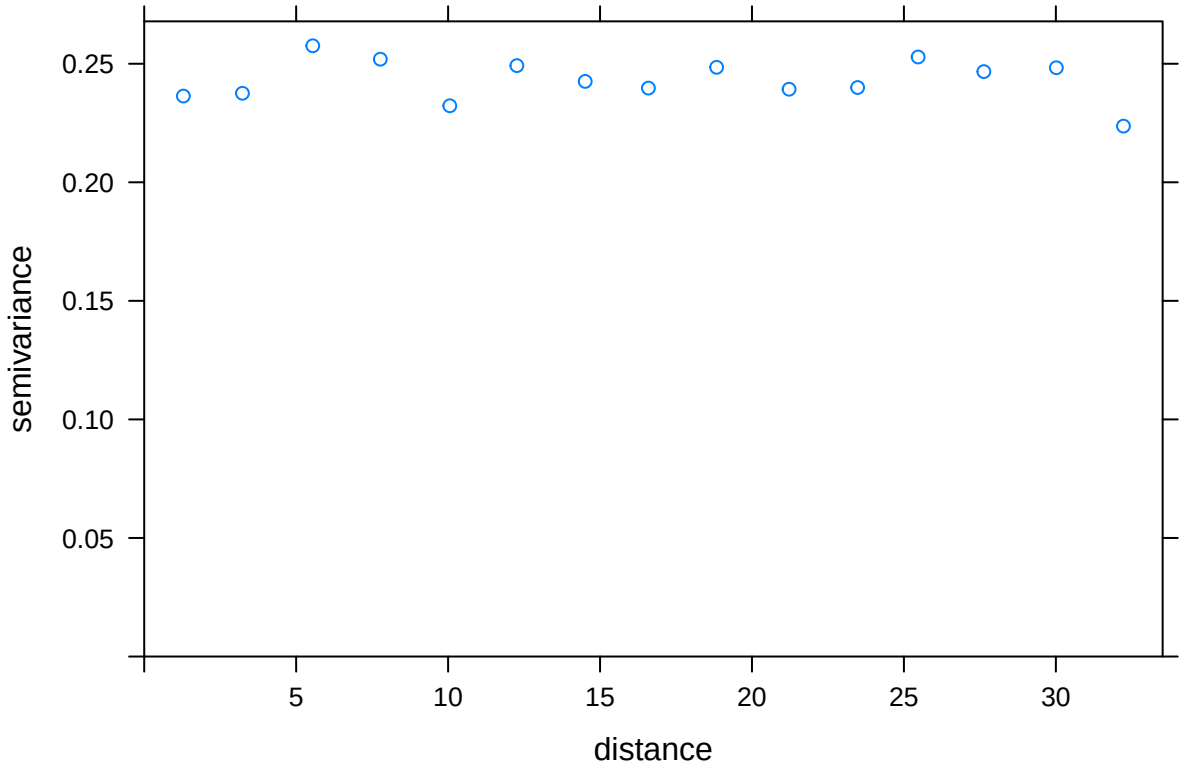
The log transformation gives the most normal distribution.

Variograms

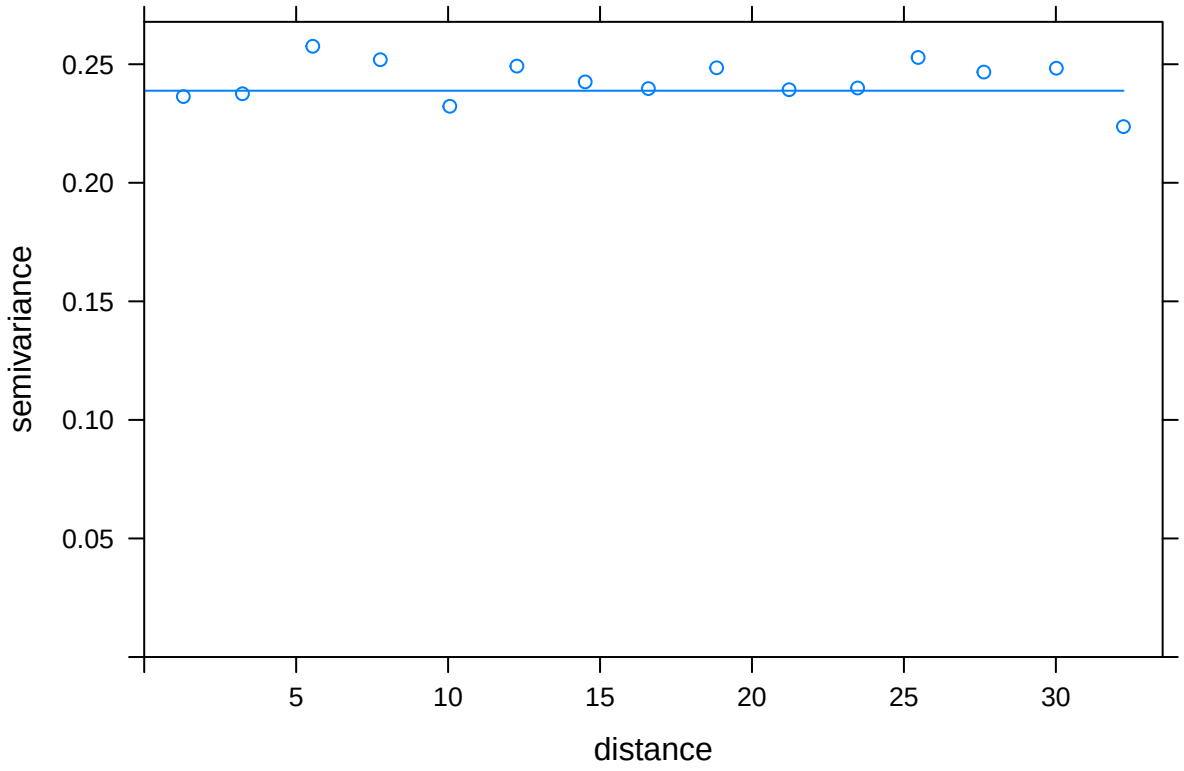
```
##        x                y              result           distance
##  Min.   : 0.7000   Min.   : 3.4000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:20.2000   1st Qu.:12.5000   1st Qu.:0.000000   1st Qu.:1.41421
##  Median :25.0000   Median :29.9000   Median :0.000000   Median :3.87298
##  Mean   :25.0389   Mean   :48.0242   Mean   :0.420382   Mean   :3.29175
##  3rd Qu.:32.8000   3rd Qu.:86.9000   3rd Qu.:1.000000   3rd Qu.:5.00000
##  Max.   :49.3000   Max.   :90.9000   Max.   :1.000000   Max.   :5.56776
##       type
##  Min.   :0.00000
##  1st Qu.:2.39790
##  Median :2.39790
##  Mean   :2.27412
##  3rd Qu.:2.48491
##  Max.   :3.09104
```

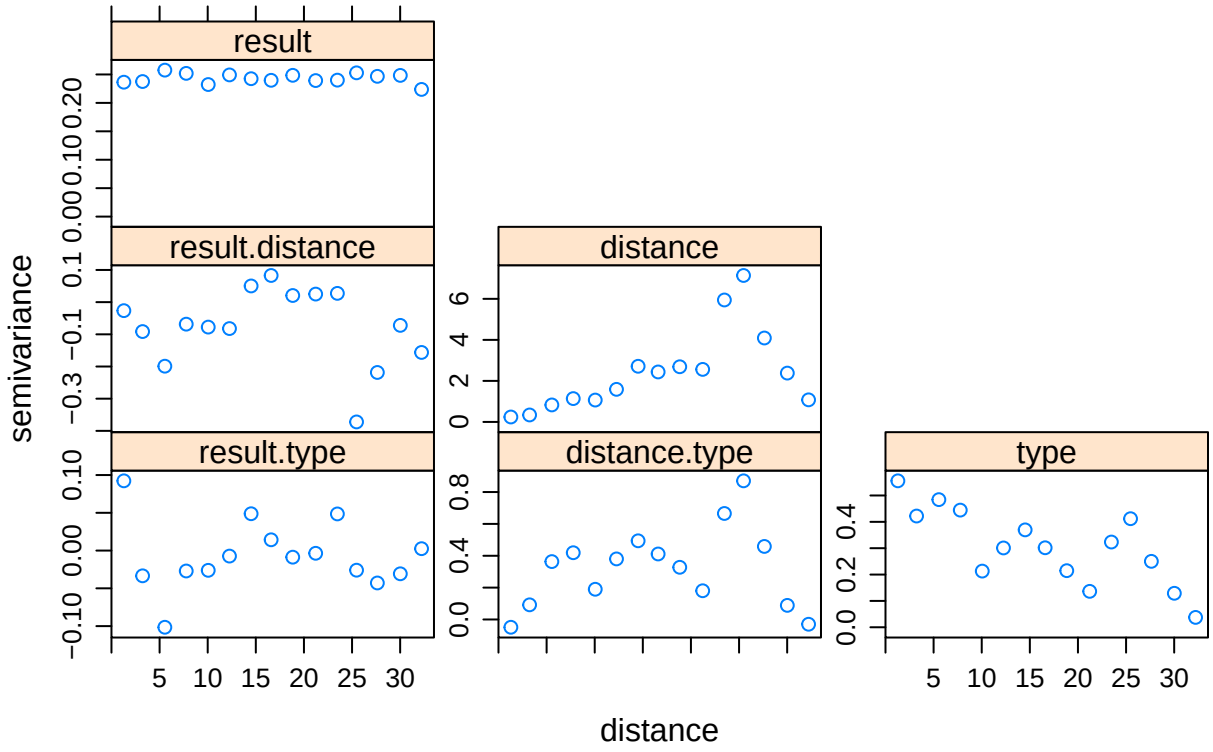
Create gstat object

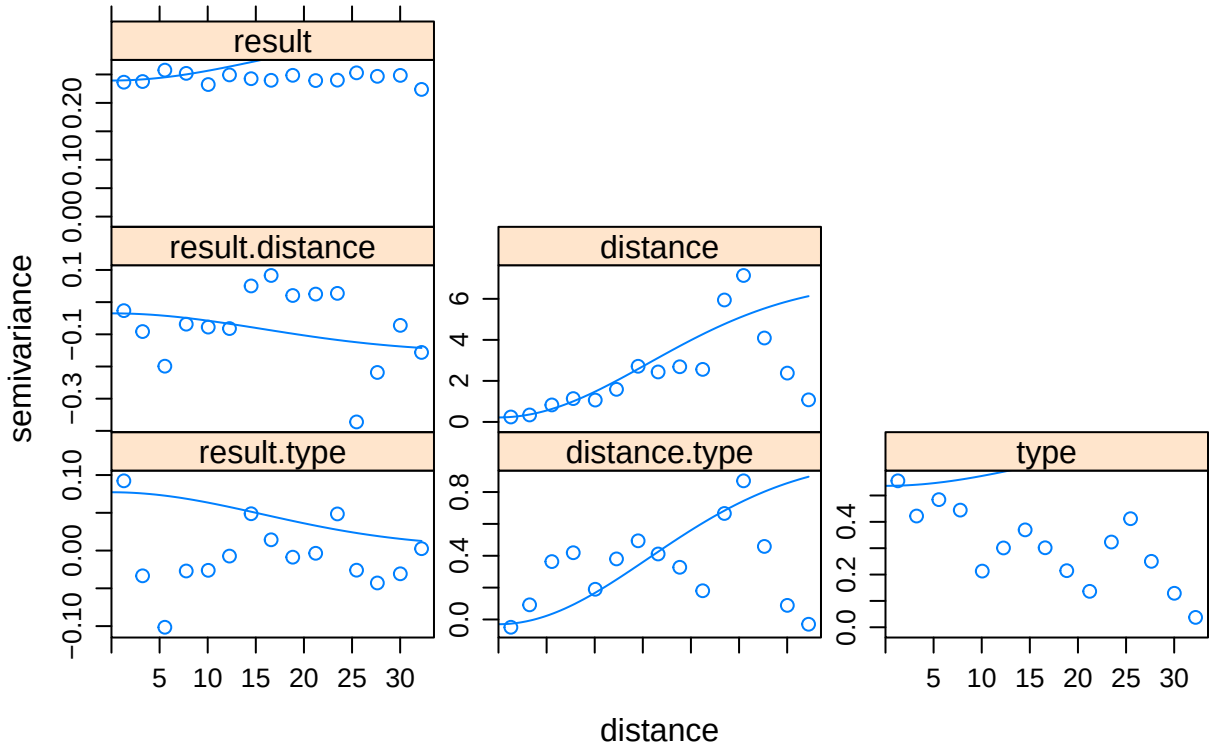



semivariance
0.25
0.20
0.15
0.10
0.05
5
10
15
20
25
30
distance

Fit a model to the variogram




result
result.distance
result.type
distance
distance.type
type
semivariance
distance
−0.10
0.00
0.10
−0.3
−0.1
0.1
0.20
0
2
4
6
0.0
0.4
0.8
0.2
0.4
5
10
15
20
25
30




```
## data:
## result : formula = result`~`1 ; data dim = 157 x 3
## distance : formula = distance`~`1 ; data dim = 157 x 3
## type : formula = type`~`1 ; data dim = 157 x 3
## variograms:
##                      model         psill range
## result[1]              Nug  0.2393296685     0
## result[2]              Gau  0.0917812141    22
## distance[1]            Nug  0.2193894480     0
## distance[2]            Gau  6.6946449885    22
## type[1]                Nug  0.5365976191     0
## type[2]                Gau  0.2005823512    22
## result.distance[1]     Nug -0.0348303475     0
## result.distance[2]     Gau -0.1216488478    22
## result.type[1]         Nug  0.0772111815     0
## result.type[2]         Gau -0.0730602258    22
## distance.type[1]       Nug -0.0297908431     0
## distance.type[2]       Gau  1.0487230659    22
## ~x + y

## [using ordinary kriging]

## [using universal kriging]

## [using ordinary kriging]
```

Cross Validation

Comparing the errors for different kriging

Ordinary Kriging

```
## [1] 38.7467949
```

Universal Kriging

```
## [1] 39.487364
```

Co-Kriging

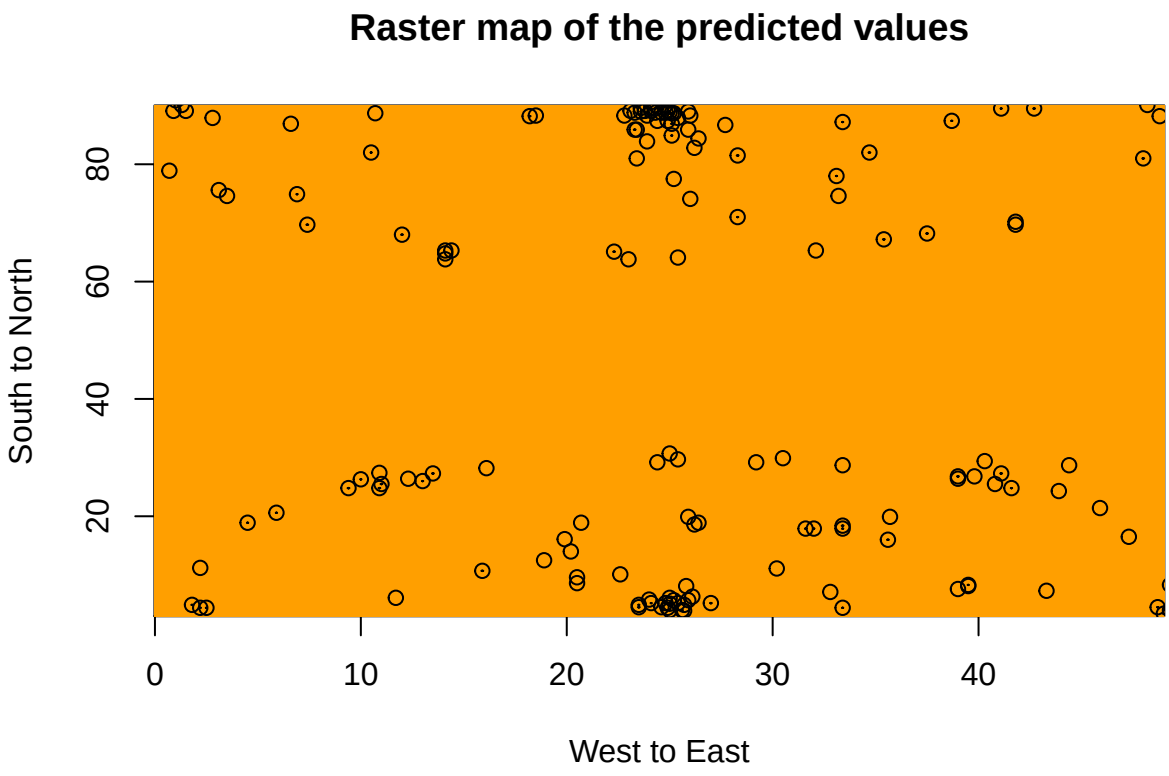
```
## [1] 38.5119539
```

Indicator Kriging

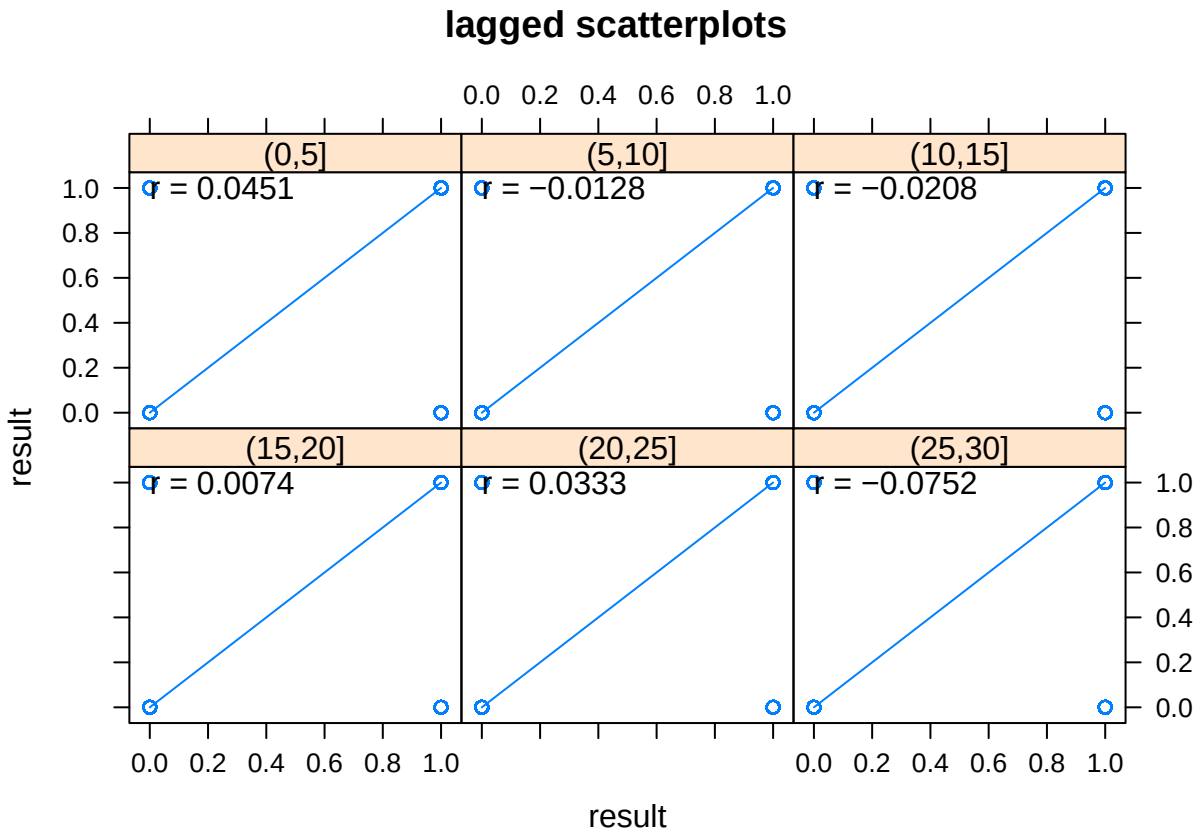
```
## [1] 38.7467949
```

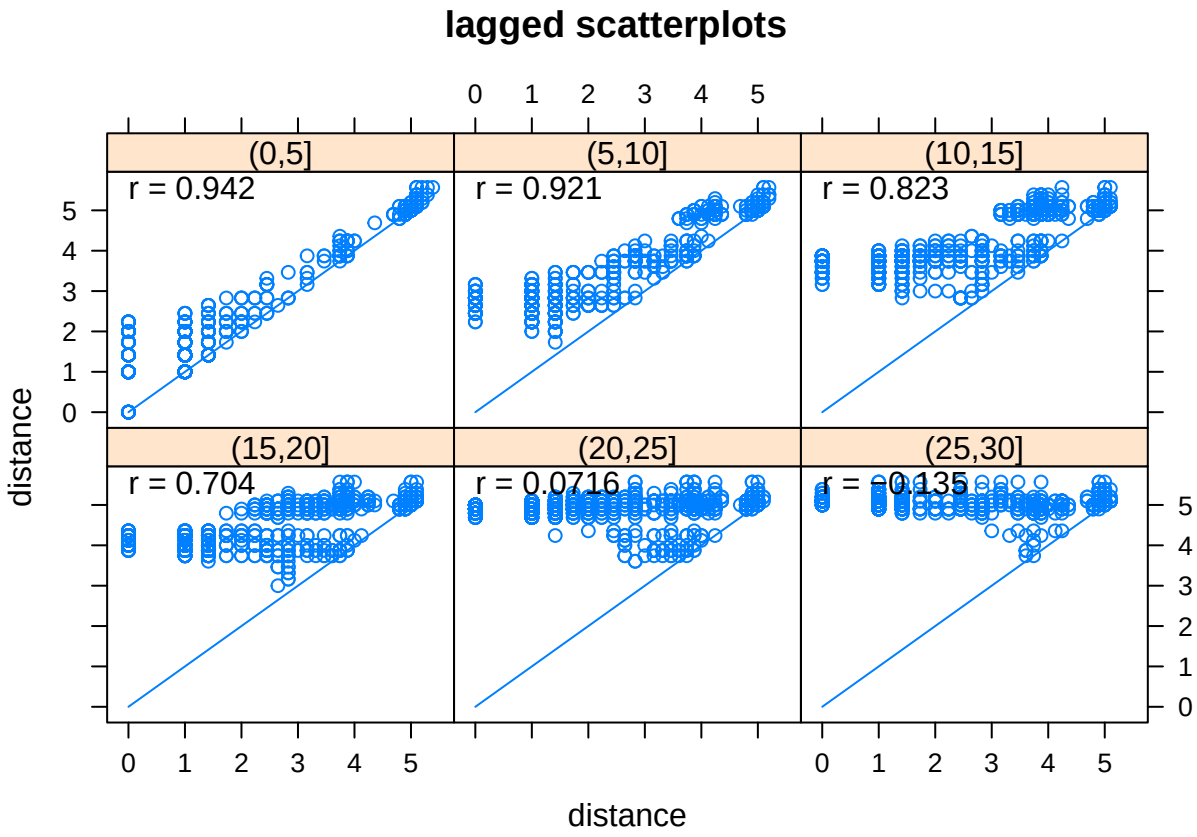
After performing cross validation and calculating the predicted residual sum of squares for each predictor, the best predictor is considered to be the indicator kriging. This is the predictor that will be used to produce the raster map.

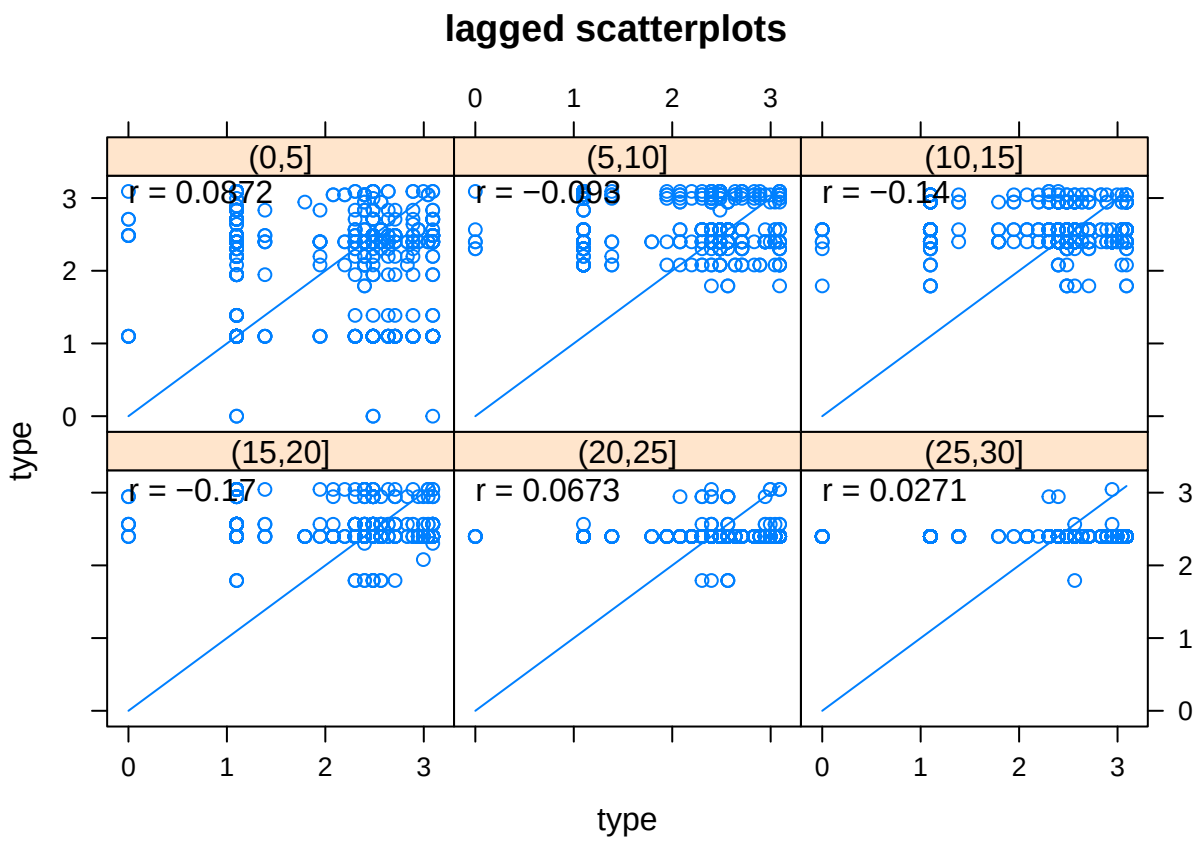
Raster Map



H-Scatterplot




lagged scatterplots
(0,5]
(5,10]
(10,15]
(15,20]
(20,25]
(25,30]
r = 0.942
r = 0.921
r = 0.823
r = 0.704
r = 0.0716
r = −0.135
distance
distance


lagged scatterplots
(0,5]
(5,10]
(10,15]
(15,20]
(20,25]
(25,30]
r = 0.0872
r = −0.093
r = −0.14
r = −0.17
r = 0.0673
r = 0.0271
type
type