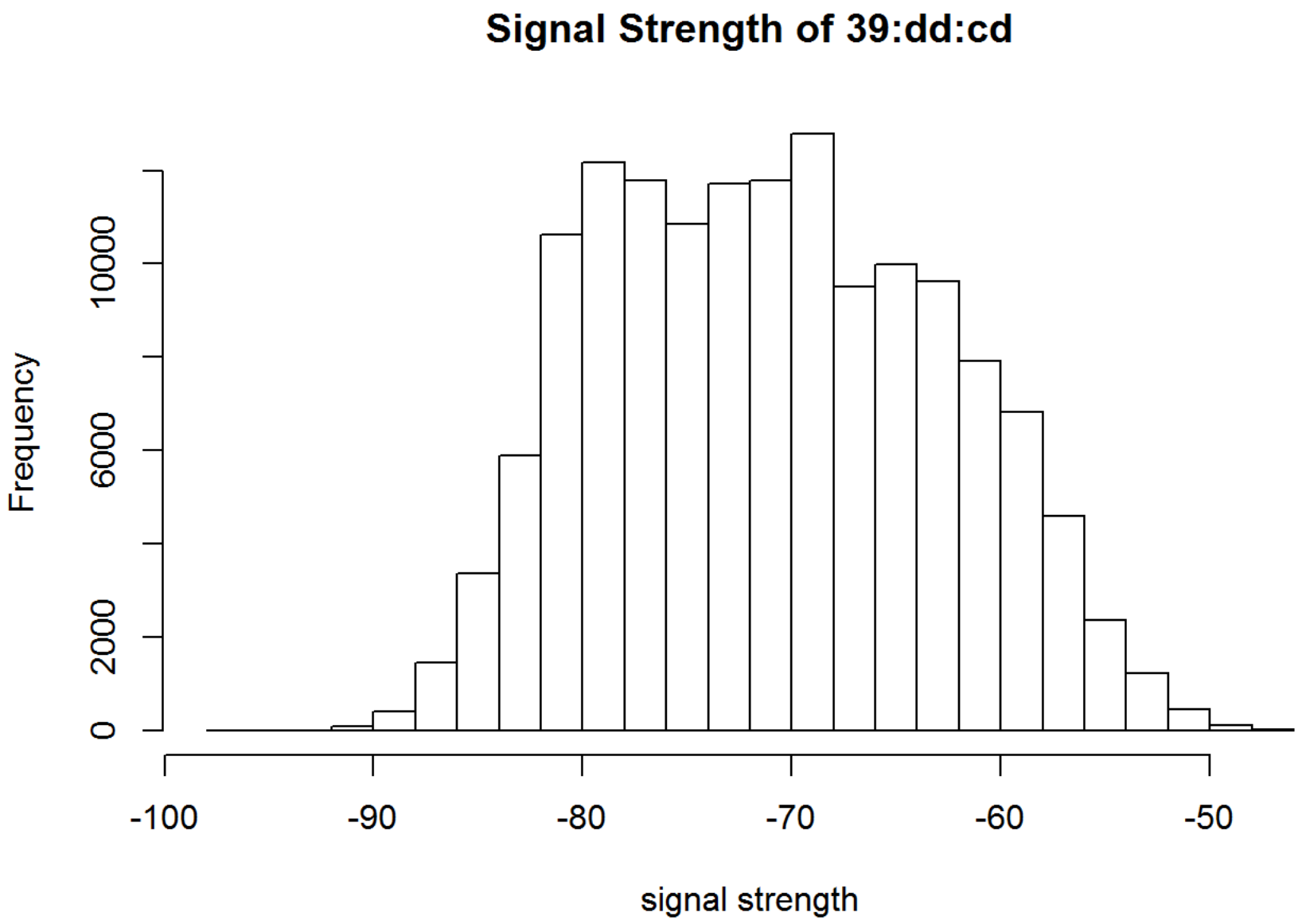
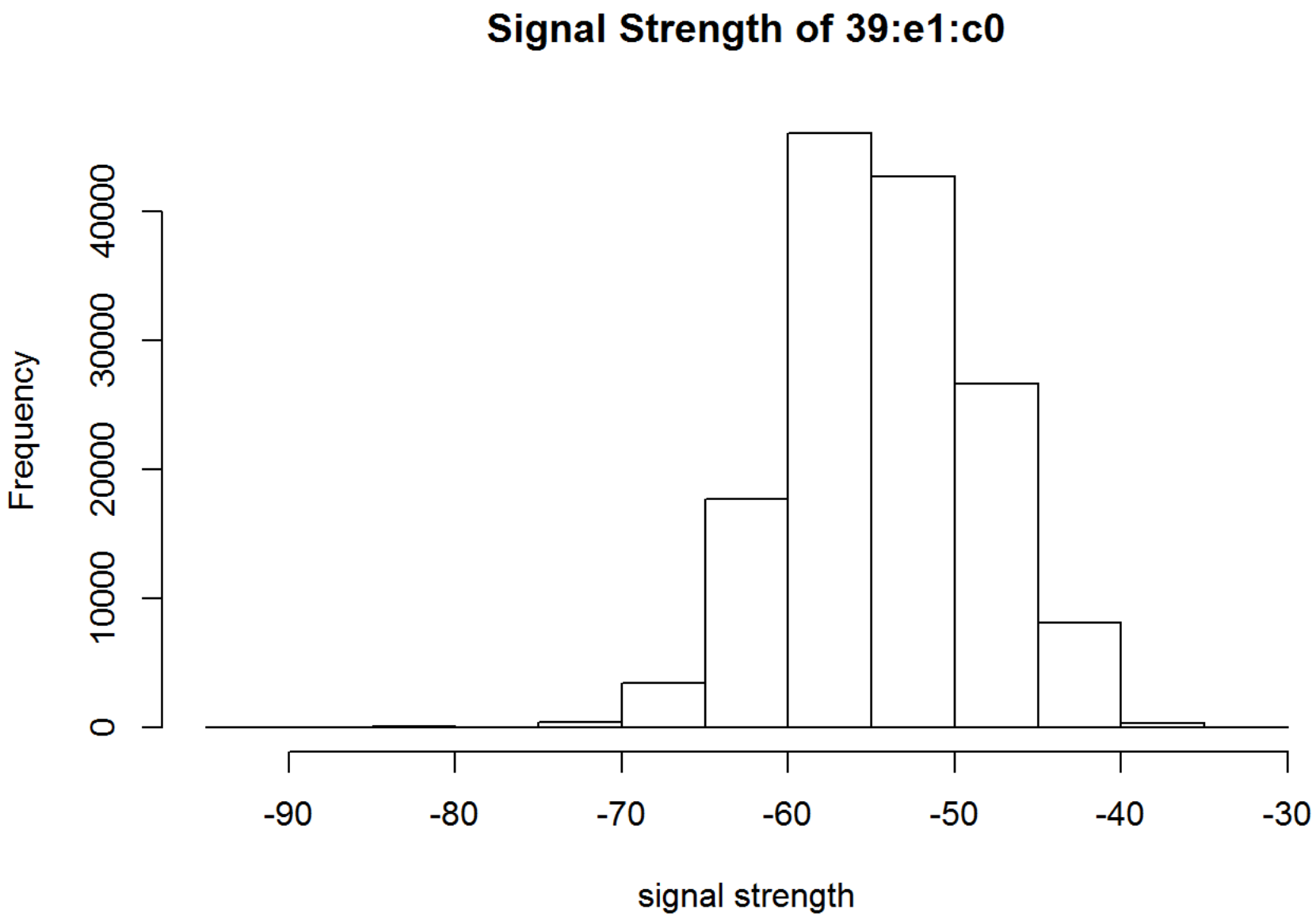
Project2 Project Group 49

```
##             time                        scanMac             posX
##  1139648662194:     21   00:02:2D:21:0F:33:1181628   1.0    :109816
##  1139647399135:     18                               2.0    :108732
##  1139648236983:     18                               0.0    : 92937
##  1139645134605:     17                               12.0   : 41960
##  1139645147006:     17                               10.0   : 41828
##  1139647640794:     17                               13.0   : 41707
##  (Other)      :1181520                               (Other):744648
##       posY          posZ          orientation
##  3.0    :241614   0.0:1181628   90.3   :  26847
##  8.0    :236069                 270.4  :  21524
##  7.0    :235221                 225.0  :  21169
##  4.0    :113511                 135.1  :  20762
##  6.0    : 85445                 315.4  :  20220
##  5.0    : 85008                 180.3  :  20183
##  (Other):184760                 (Other):1050923
##                 mac              signal             channel
##  00:0f:a3:39:e1:c0:145862   -59    : 40878   2462000000:189774
##  00:0f:a3:39:dd:cd:145619   -62    : 39630   2437000000:152124
##  00:14:bf:b1:97:8a:132962   -58    : 39628   2412000000:145619
##  00:14:bf:3b:c7:c6:126529   -60    : 38970   2432000000:126529
##  00:14:bf:b1:97:90:122315   -57    : 38689   2427000000:122315
##  00:14:bf:b1:97:8d:121325   -63    : 38198   2442000000:121325
##  (Other)          :387016   (Other):945635   (Other)   :323942
##  type
##  1:203185
##  3:978443
```

```
##
##
##
##
##
```

```
## 00:04:0e:5c:23:fc 00:0f:a3:39:dd:cd 00:0f:a3:39:e0:4b 00:0f:a3:39:e1:c0
##               418            145619             43508            145862
## 00:0f:a3:39:e2:10 00:14:bf:3b:c7:c6 00:14:bf:b1:97:81 00:14:bf:b1:97:8a
##             19162            126529            120339            132962
## 00:14:bf:b1:97:8d 00:14:bf:b1:97:90 00:30:bd:f8:7f:c5 00:e0:63:82:8b:a9
##            121325            122315               301               103
## 02:00:42:55:31:00 02:0a:3d:06:94:88 02:2e:58:22:f1:ac 02:37:fd:3b:54:b5
##            103887                 1             25112              8732
## 02:42:1c:4e:b5:c0 02:4f:99:43:30:cd 02:5c:e0:50:49:de 02:64:fb:68:52:e6
##                 1                 1              6997             50852
## 02:b7:00:bb:a9:35
##              7602
```






```
## 2412000000 2422000000 2427000000 2432000000 2437000000 2442000000
##          0          0          0          0          0          0
## 2447000000 2457000000 2462000000 2472000000
##          0          0     145862          0
```

```
## 2412000000 2422000000 2427000000 2432000000 2437000000 2442000000
##           0          0          0     126529          0          0
## 2447000000 2457000000 2462000000 2472000000
##           0          0          0          0
```

```
## 2412000000 2422000000 2427000000 2432000000 2437000000 2442000000
##           0     120339          0          0          0          0
## 2447000000 2457000000 2462000000 2472000000
##           0          0          0          0
```

```
## 2412000000 2422000000 2427000000 2432000000 2437000000 2442000000
##           0          0          0          0     132962          0
## 2447000000 2457000000 2462000000 2472000000
##           0          0          0          0
```

```
## 2412000000 2422000000 2427000000 2432000000 2437000000 2442000000
##           0          0          0          0          0     121325
## 2447000000 2457000000 2462000000 2472000000
##           0          0          0          0
```

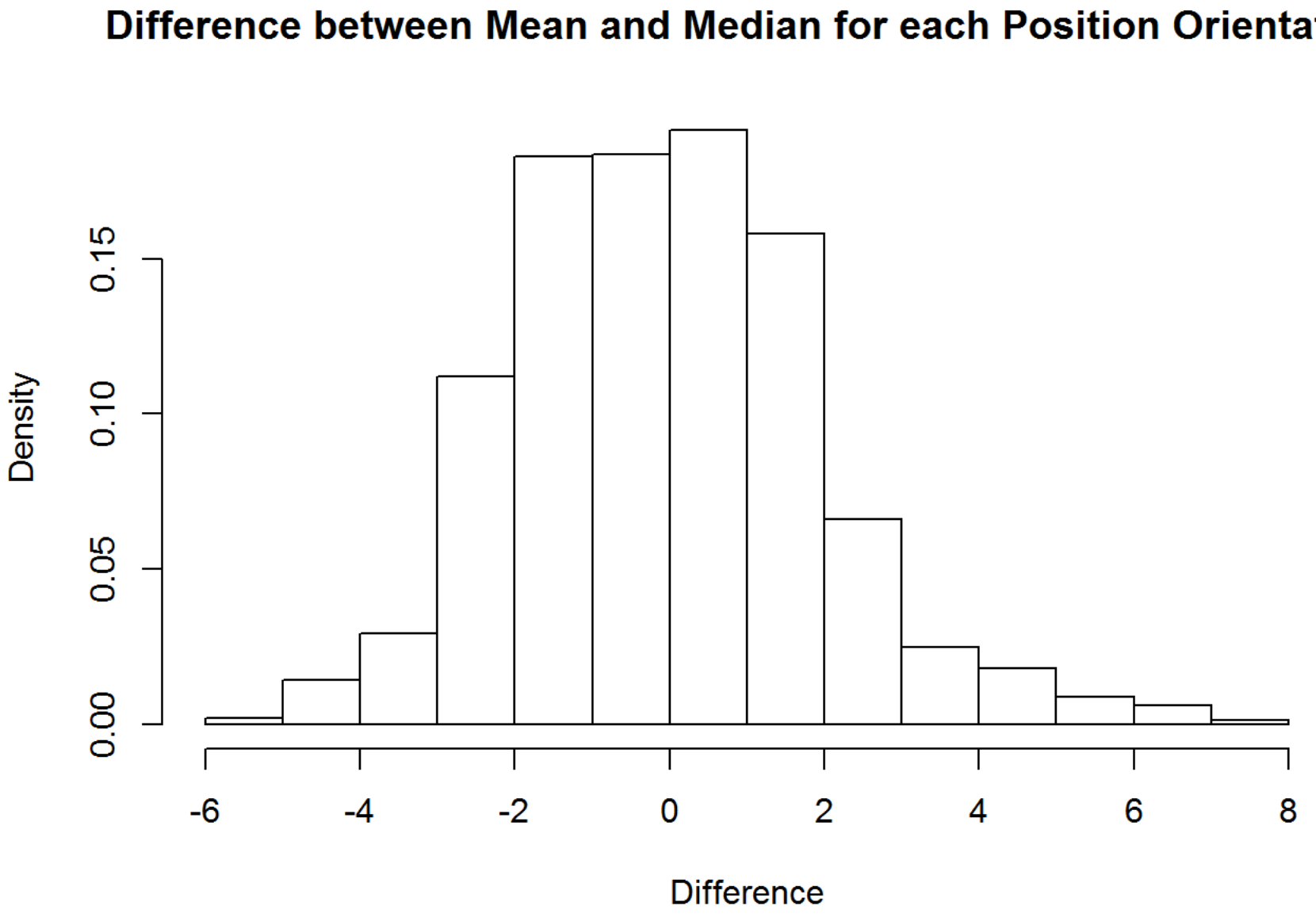
```
## 2412000000 2422000000 2427000000 2432000000 2437000000 2442000000
##           0          0     122315          0          0          0
## 2447000000 2457000000 2462000000 2472000000
##           0          0          0          0
```

From the above summary, we can see that we only have one scanMac data which is just the hand-held device, there is no need to keep this information. The posZ is also unique for all data, so we drop it. And according to the description, type

specifies whether the device is adhoc or not. So we drop the data with type = 1. From the number of records of the mac address, we can see that the five starting with “00:14:bf” are Cisco ones, and we’re going to keep these. And from the rest of the data, only two have about the same level of number of records. They are “00:0f:a3:39:dd:cd” and “00:0f:a3:39:e1:c0”. Then from the above two histograms of their signal strength we can see that the first one has obviously weaker strength than the second one which means that it is getting the signal maybe from another floor instead of the one we’re interested in on this floor, so we drop the first one. From the summary of the channels, we see that each mac address corresponds to only one channel. So there is no need to keep this information.

PART II

In this part, we have about 110 signals for each position orientation combination. In order to do the cross validation in a simpler way, we want to collapse these records into one. So we would want to take the mean.



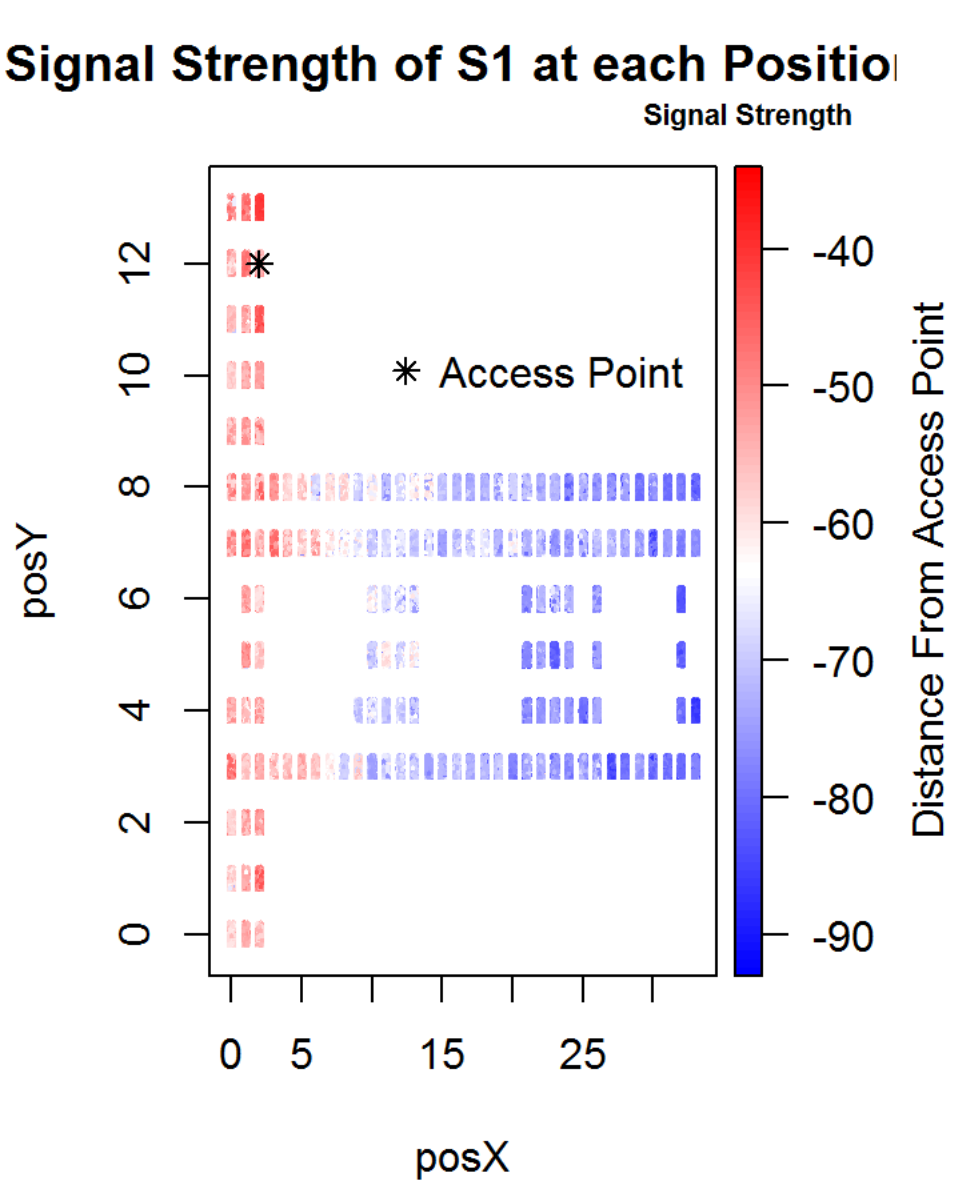
From the above graph, we can see that the mean and median are generally pretty close to each other. Because the signal strengths are roughly symmetric, we can use the mean to represent the records.

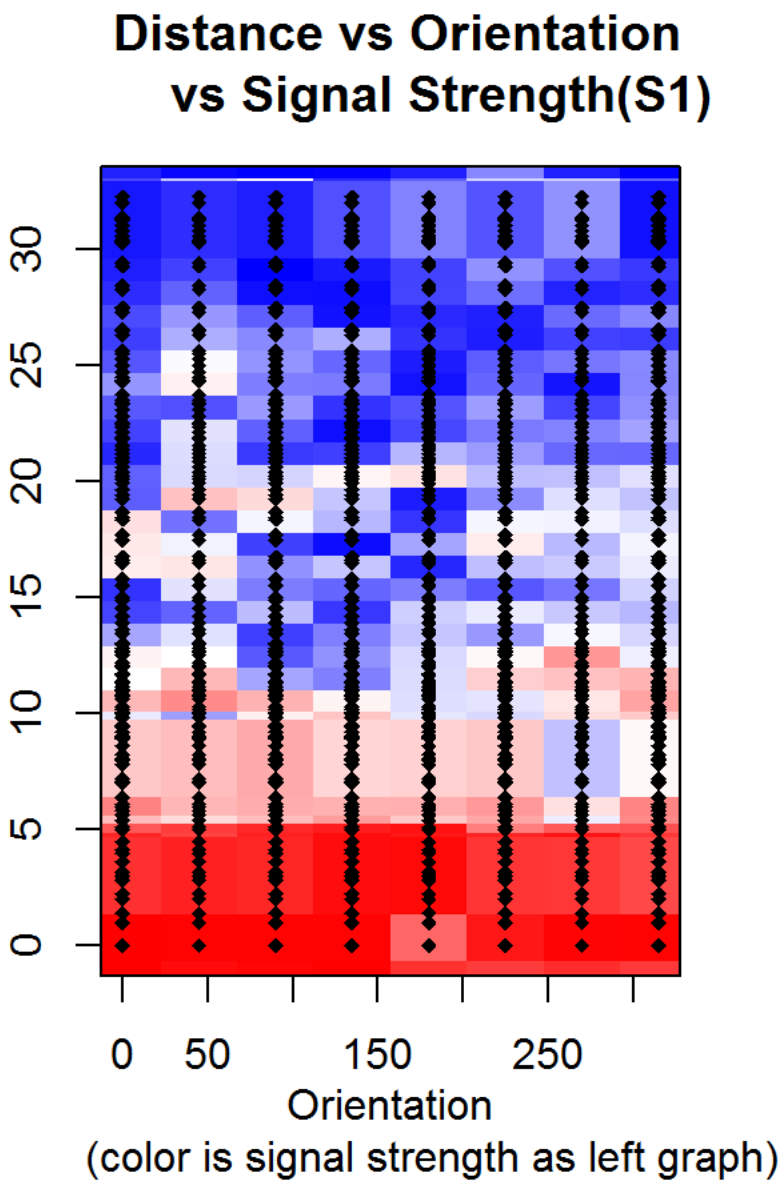
```
## [[1]]
```

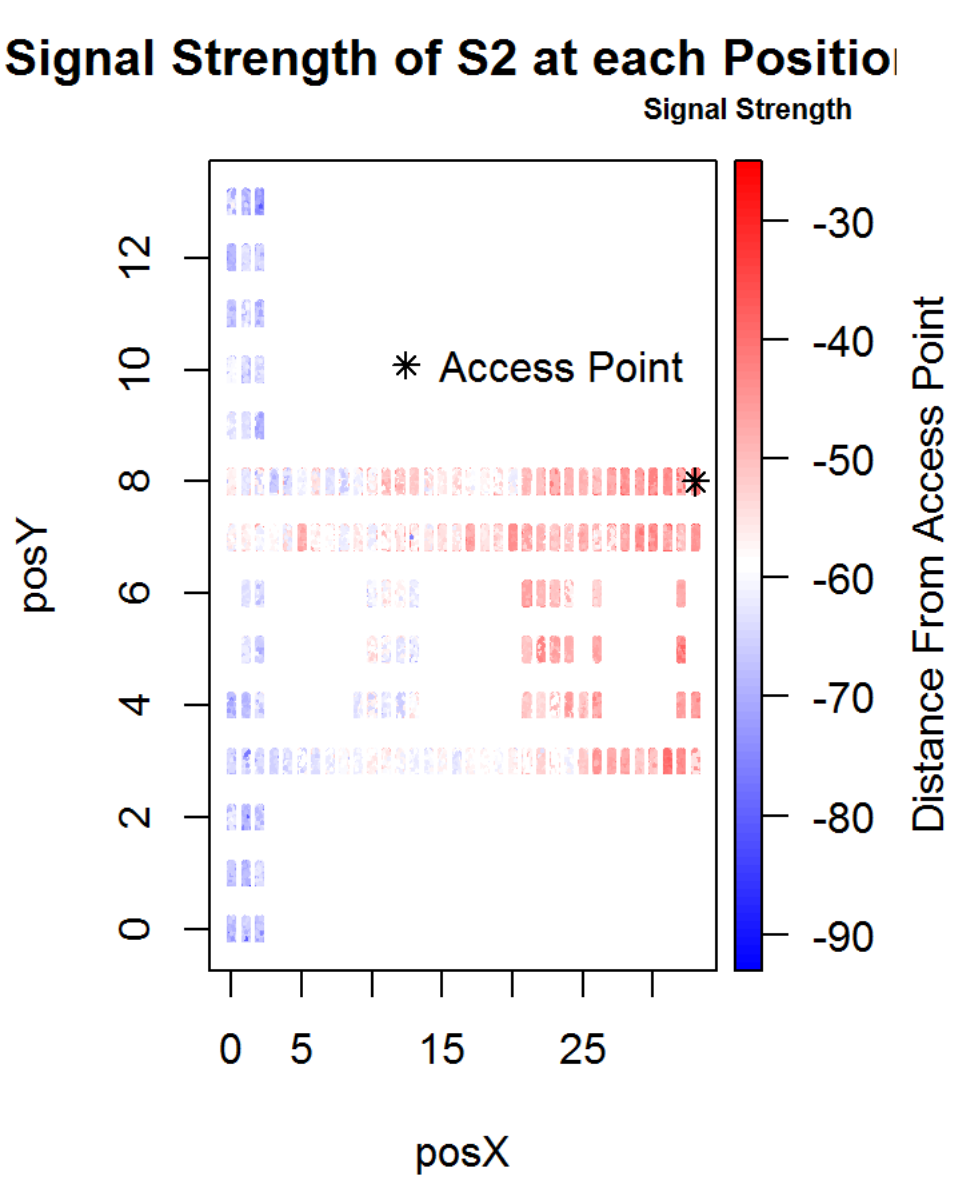
```
##    posX posY
## 15    2   12
##
## [[2]]
##     posX posY
## 489   33    8
##
## [[3]]
##     posX posY
## 833    2    0
##
## [[4]]
##     posX posY
## 734   33    4
##
## [[5]]
##      posX posY
## 1022   12    3
##
## [[6]]
##      posX posY
## 1324    8    8
```

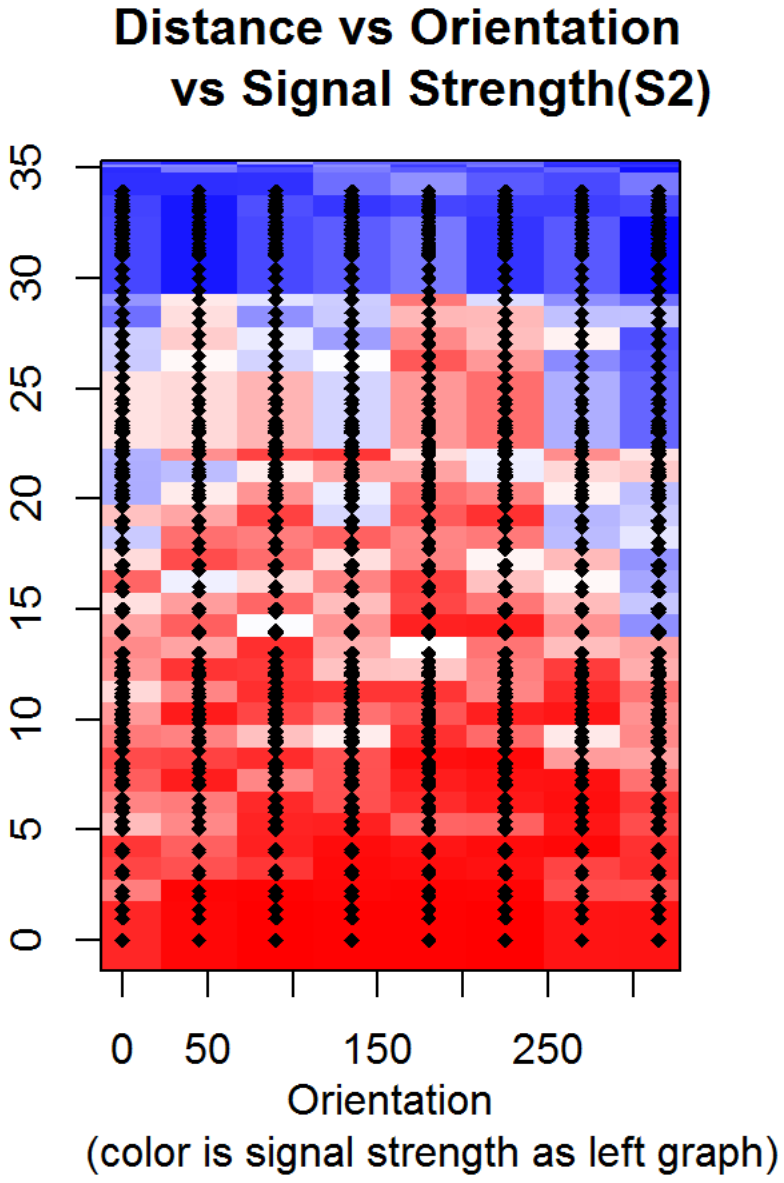
We check the max signal strength of these signal strengths, which we'll use as the position of the corresponding devices.

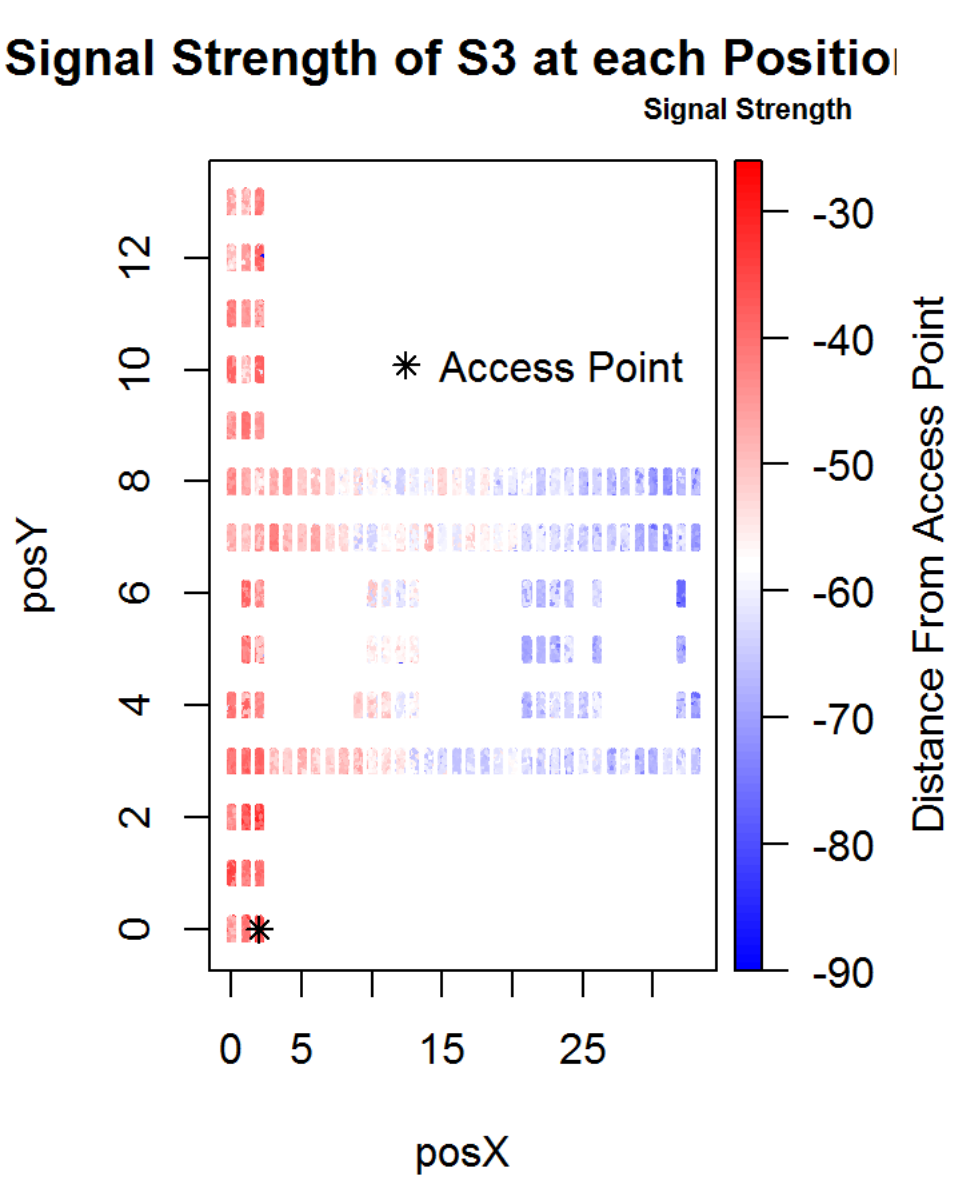
```
##
## Attaching package: 'gplots'
##
## The following object is masked from 'package:stats':
##
##     lowess
```

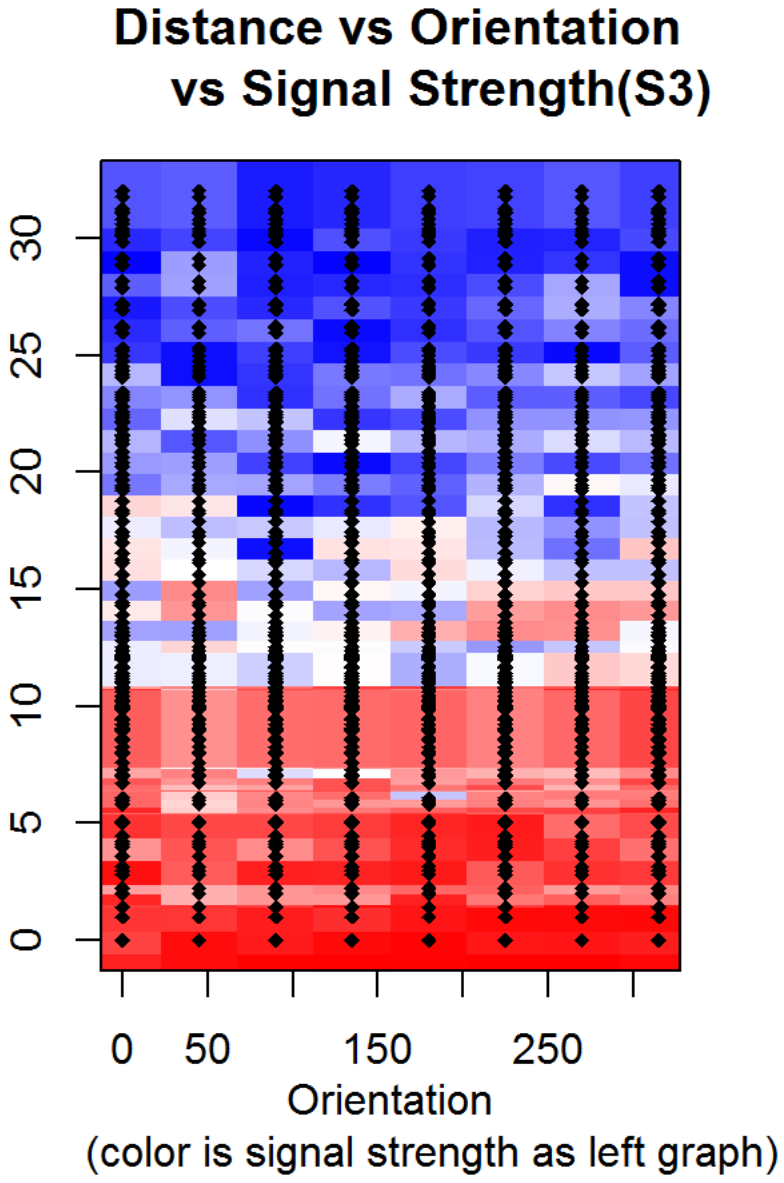


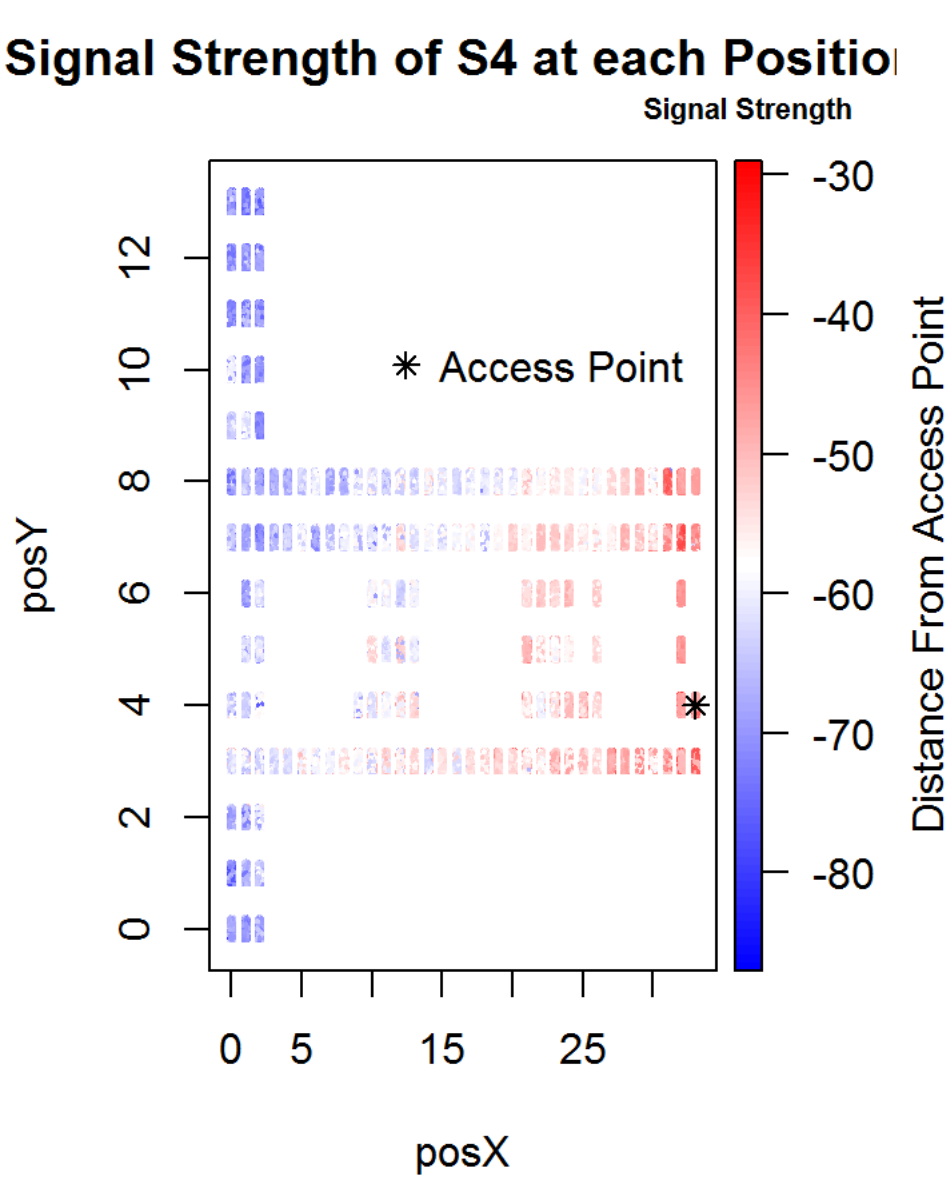


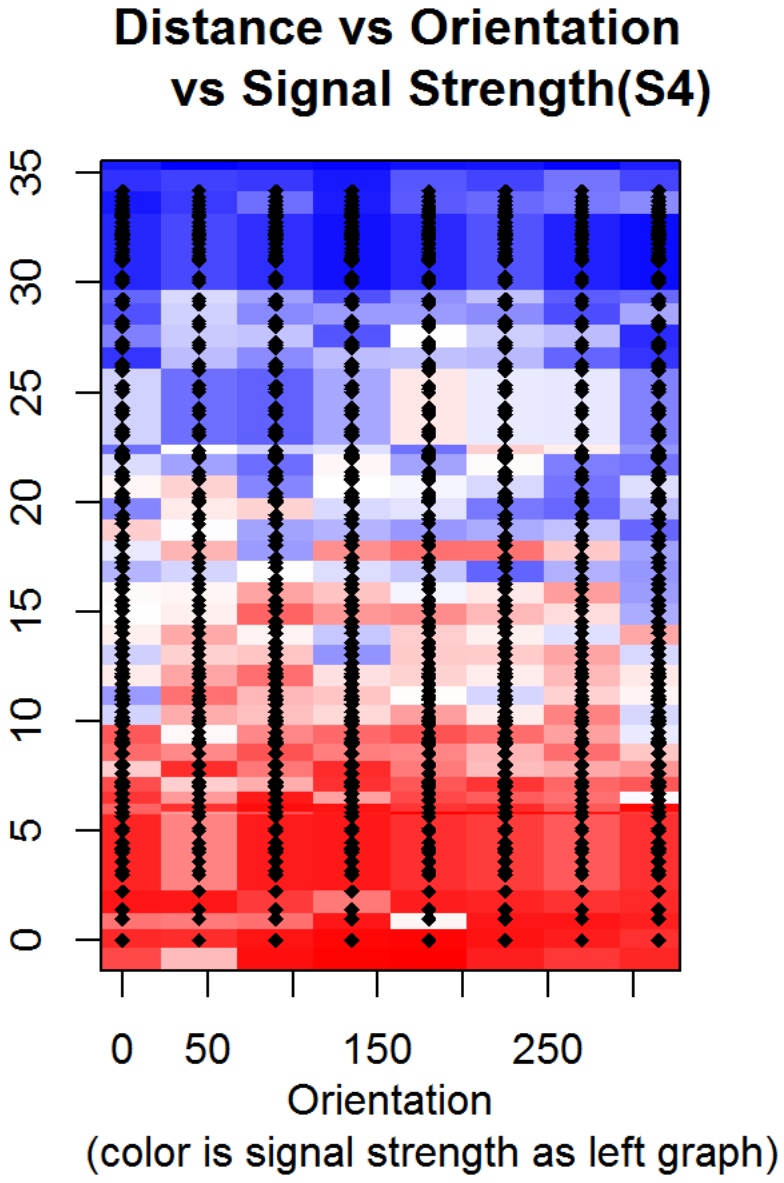


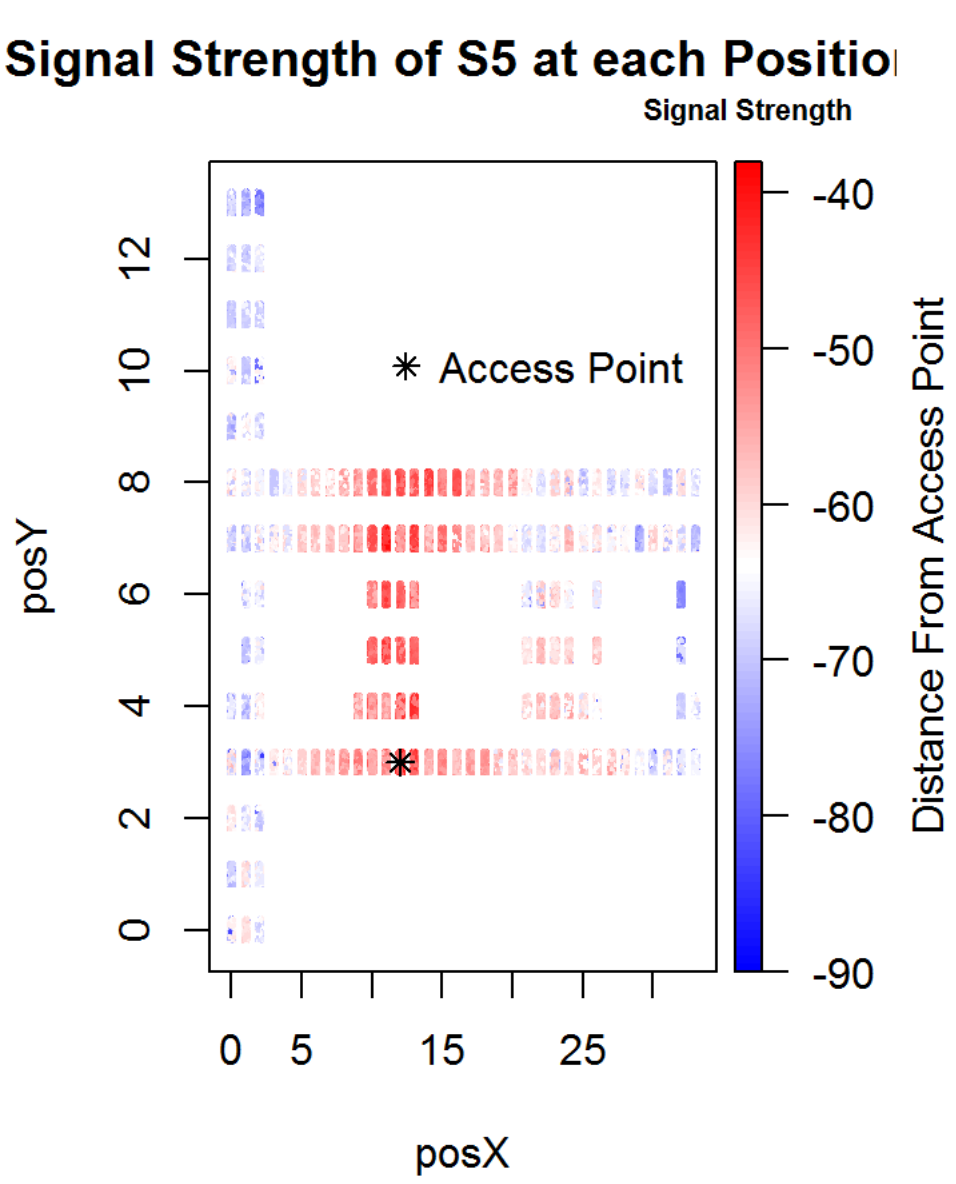


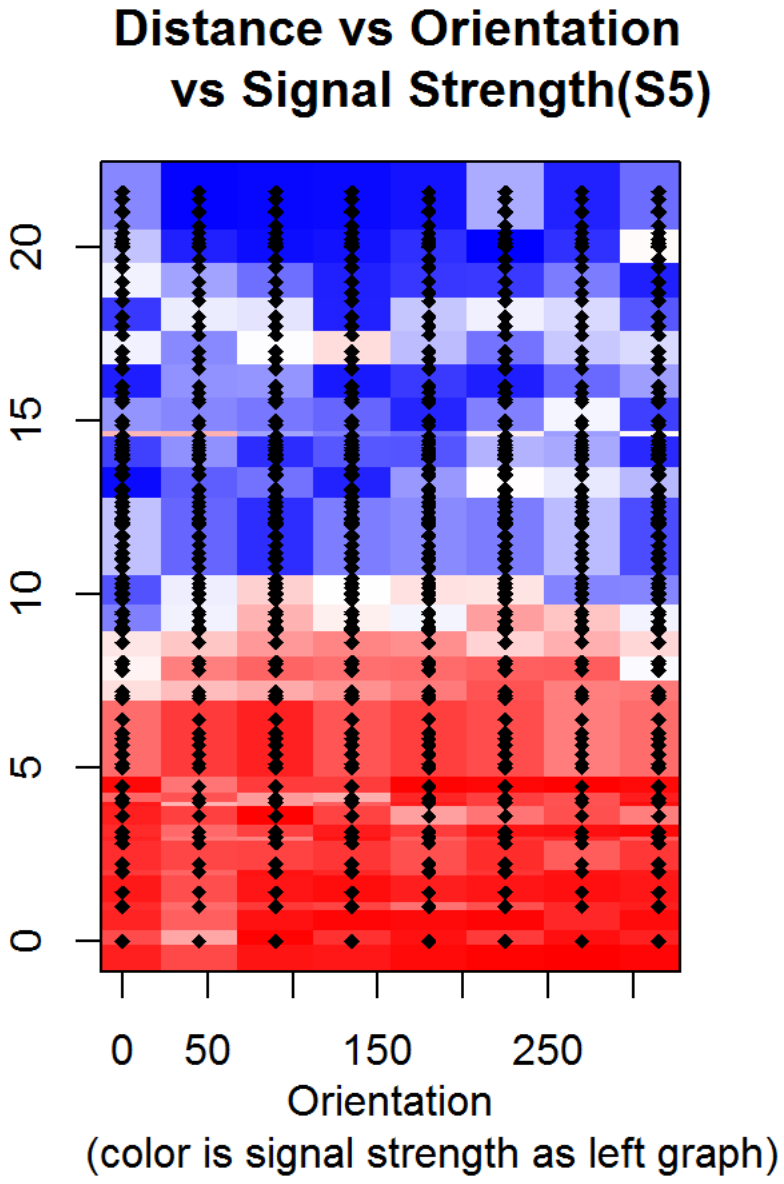


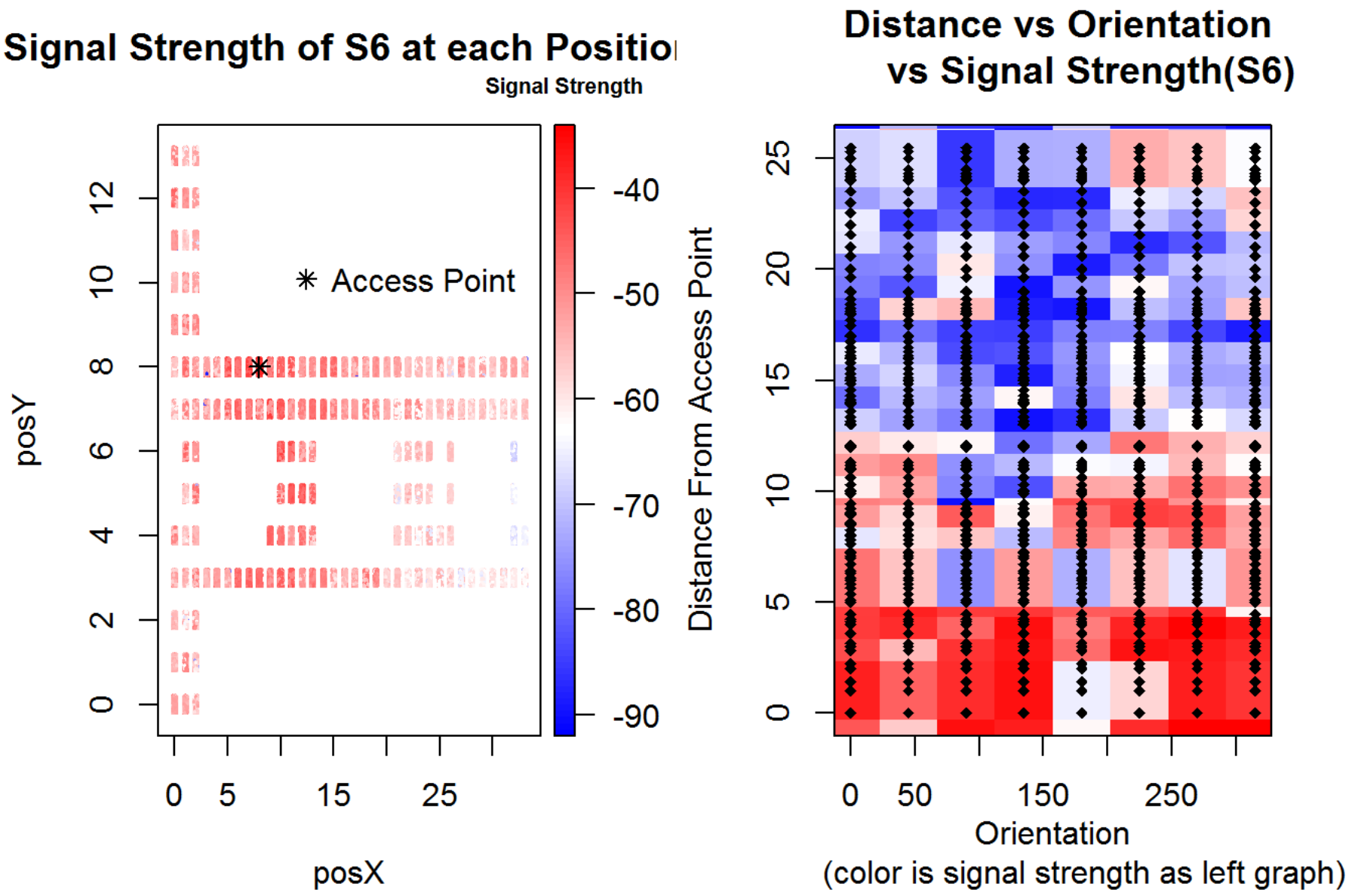












These plots show two methods of displaying signal strength readings. The left plot shows the spatial distribution of the readings, giving information about where the signal is strongest and weakest in the building.

The right plot is effective at showing how signal strength decays over distance. One aspect of the plot of note is how there is often a clear horizontal cutoff for the bright red (best signal strength) regions. This should represent the effective range of

the signal, before it begins to decay significantly. Also of note, these plots show that, generally, orientation does not significantly change signal strength readings. And we can also see from the left plot that, when there is a wall, the signal strength decays faster(i.e. it turns blue faster).

PART III

Given all the data we have now, we can do the cross validation. We first transform the data frame into the distance matrix as we had in the emails case.

The predNeighbor and cvKnn function are similar to those we used in the email case except that we modified them to work for signals. We will use the k-closest positions with corresponding orientations and take mean over their x and y values to predict the position of the one we're testing on.

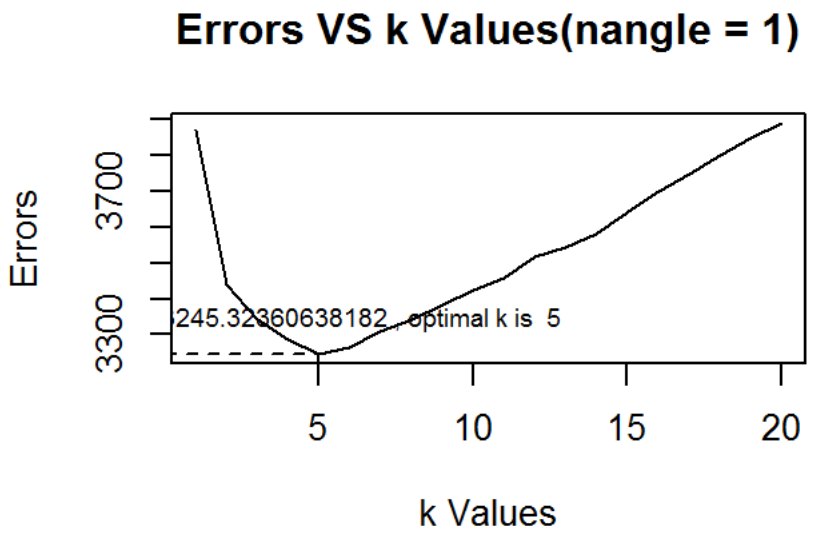
We also add angle and nangles as inputs to decide angles of the signal and decide the number of angles we are going to use for predicting positions. That is, the angles are the orientation of the positions we are predicting and nangles is the number of angles to use to predict each position(say if angle of position 1 is 45 and nangles = 3, we will use 0, 45, 90 degrees data to predict its position.)

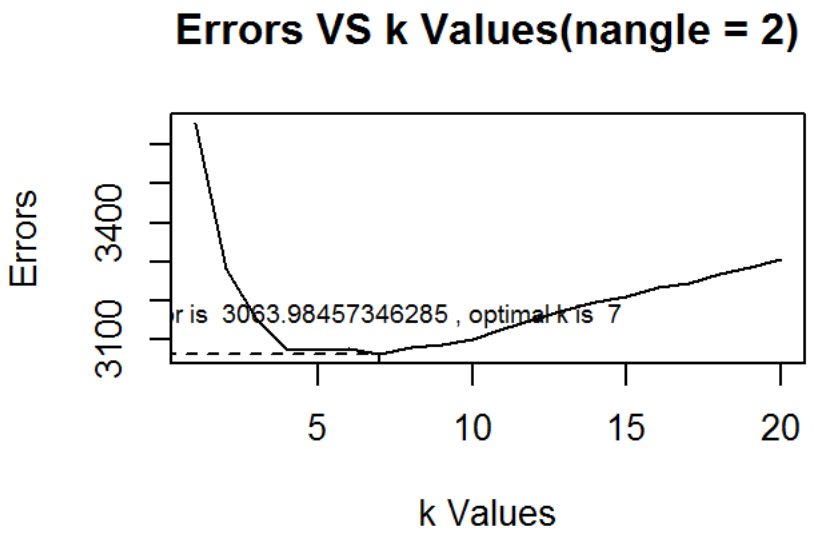
In this case, since we don't have that much data as the email case, we will use 5 folds in cvKnn instead of 10. And still we'll try k from 1 to 20.

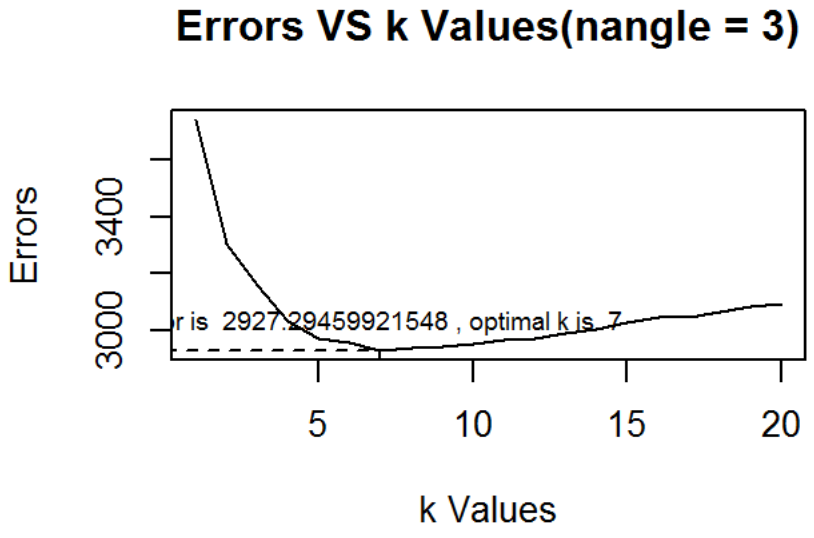
```
## [1] 5
```

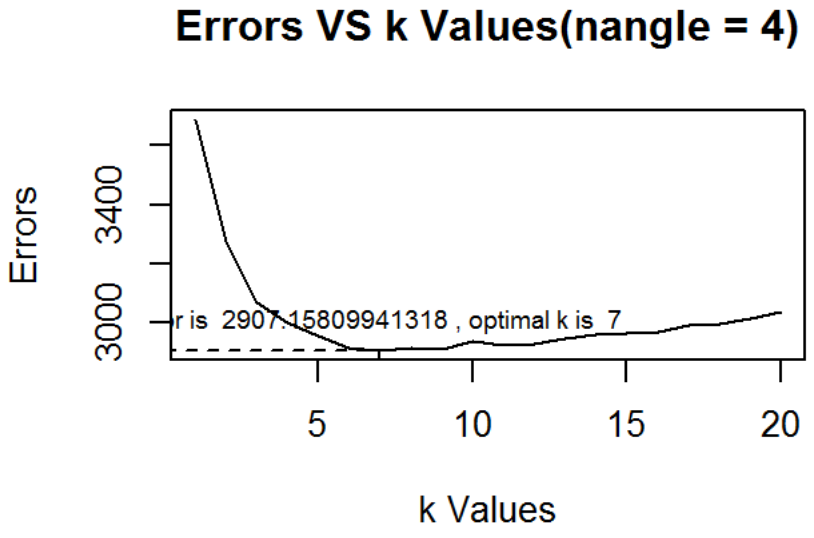
```
## [1] 7
```

```
## [1] 7
```







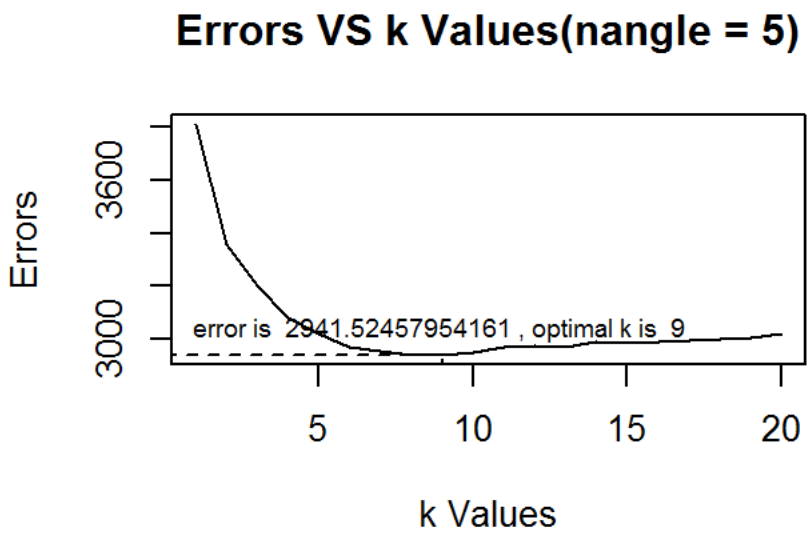


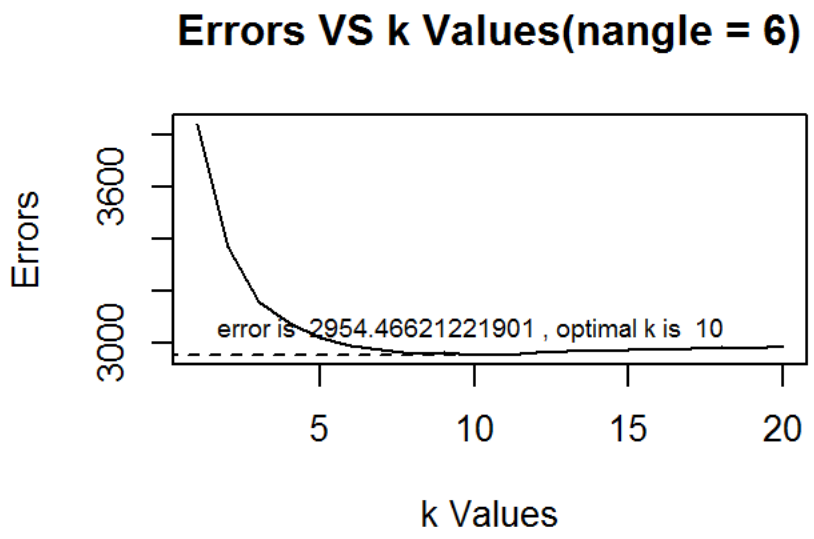

```
## [1] 7
```

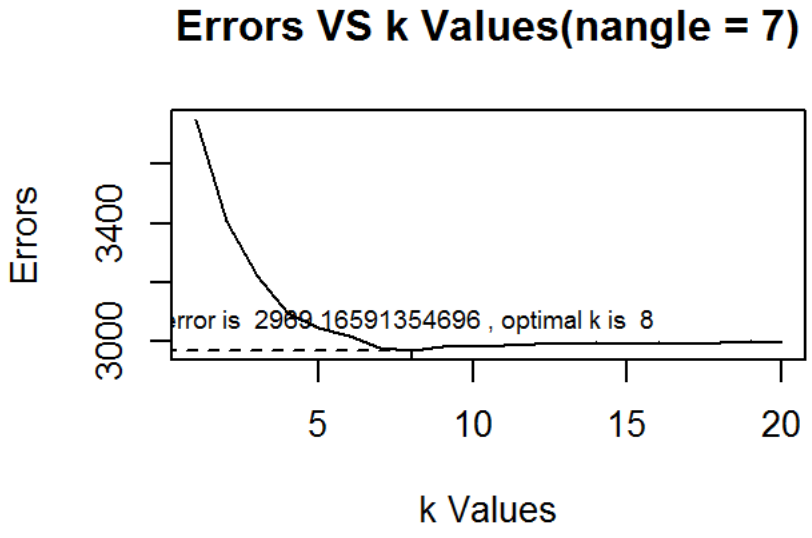
```
## [1] 9
```

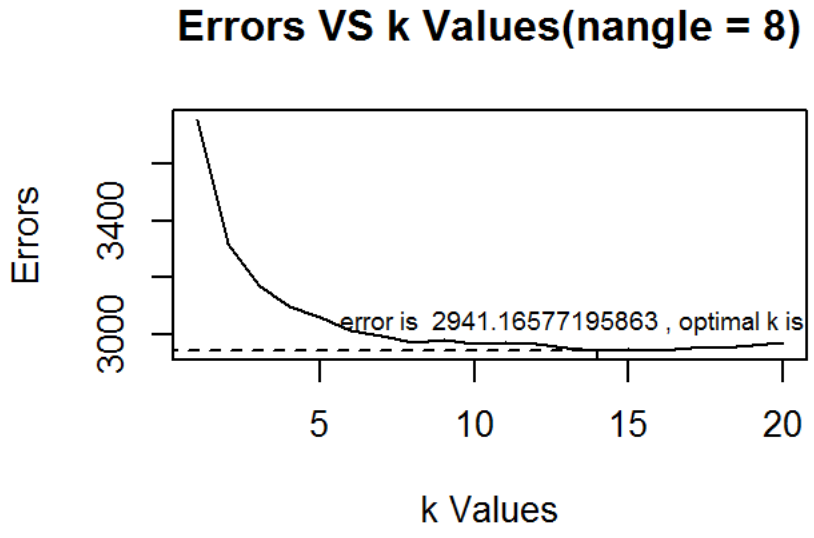
```
## [1] 10
```

```
## [1] 8
```

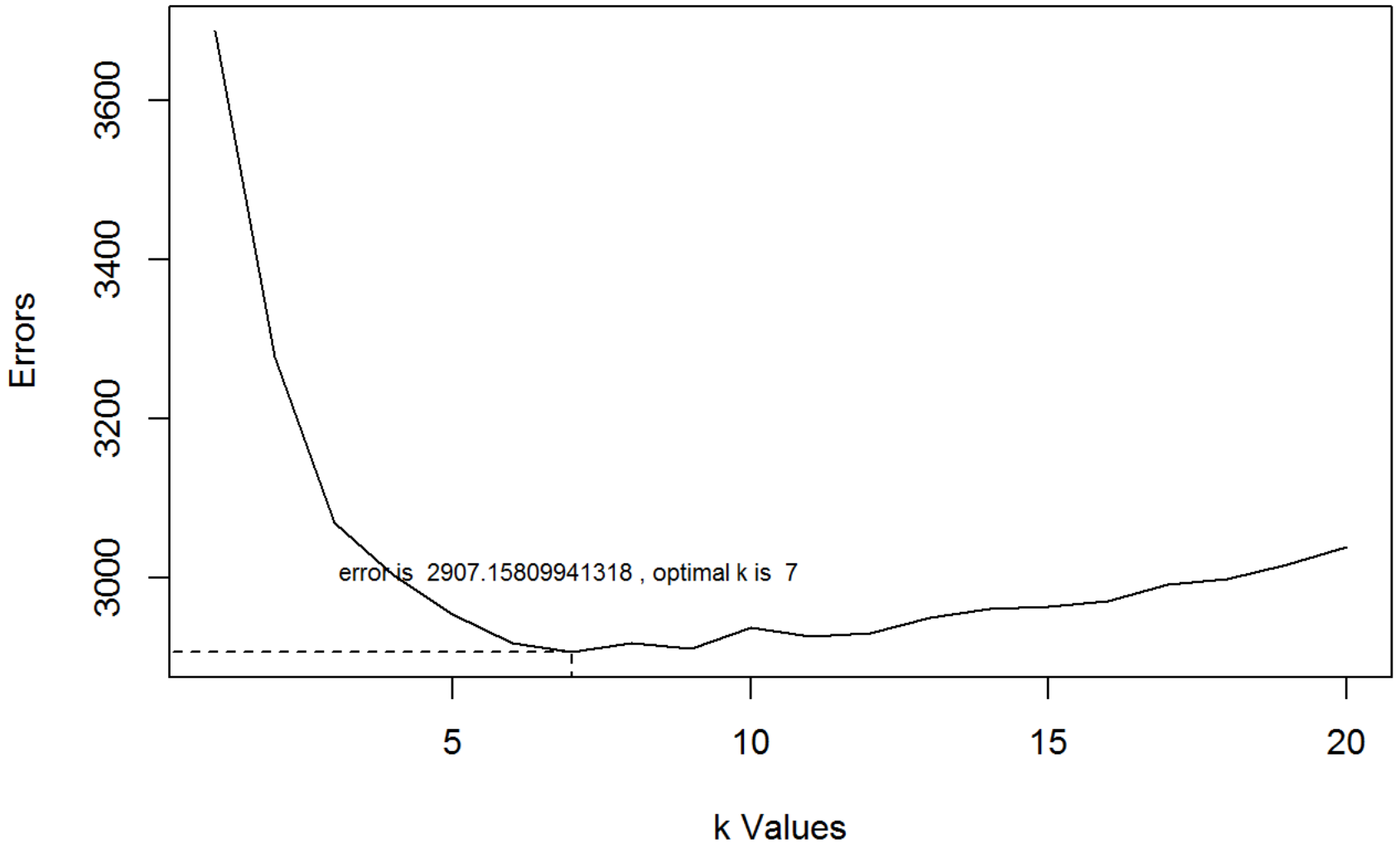








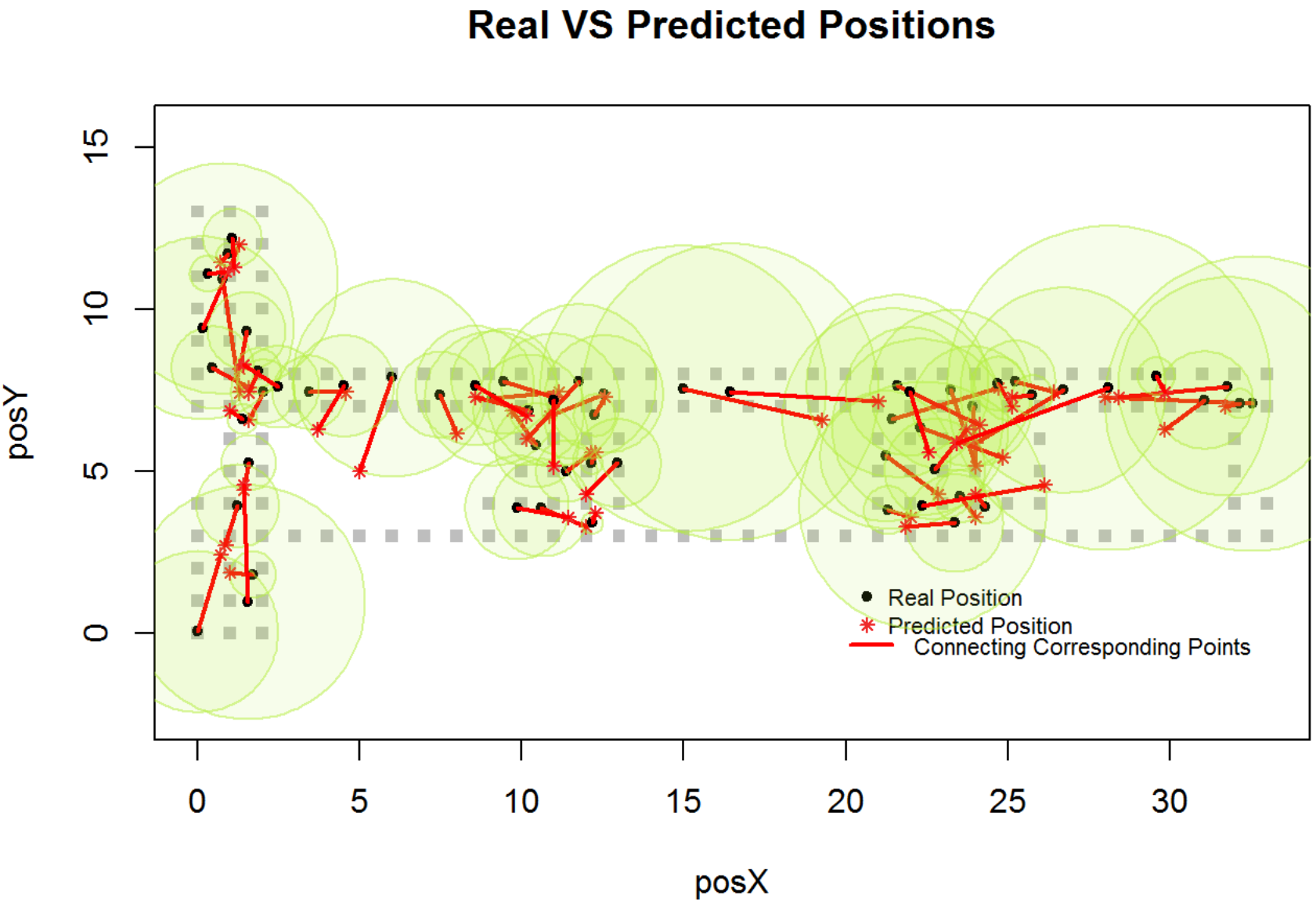

```
## [1] 14
```



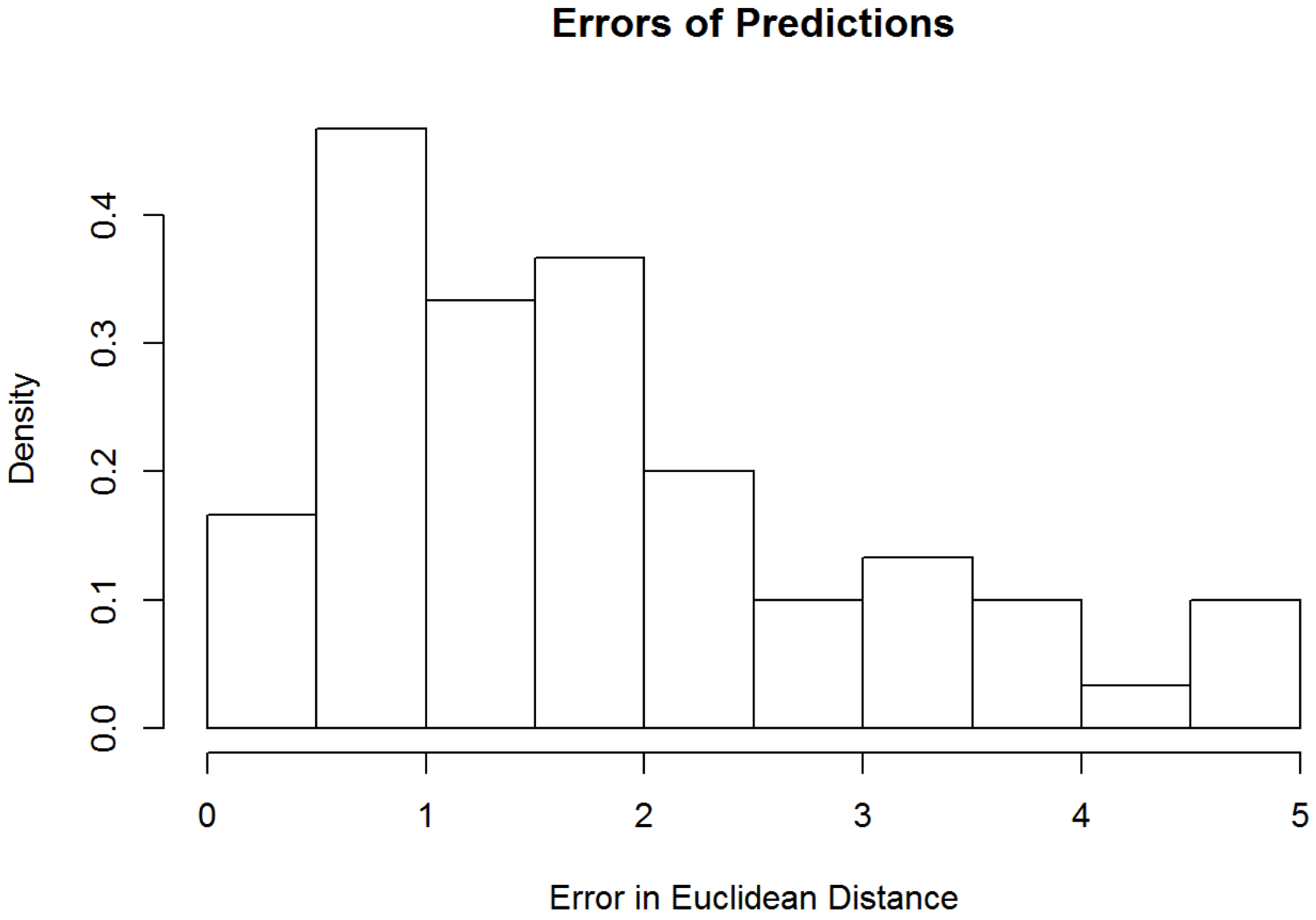
From the above graph, we can see that nangle = 4 gives us the best estimation, so we'll use this as the optimal k, which is 7 and the min error is 2907.

Then we read in the online data, clean the data and predict the positions for these data as we did for the offline data.

```
##
## Attaching package: 'plotrix'
##
## The following object is masked from 'package:gplots':
##
##     plotCI
```



Here, we see the spatial distribution of both real and predicted access point positions. The black point shows the real position, the red point shows the predicted position, and a light circle overlay has been added with the radius of the Euclidian distance from real to predicted points in order to give a sense of the overall error. From the graph, we can see that generally, the error is not large.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3316  0.8409  1.5060  1.8000  2.4780  4.9940
```

This histogram corresponds to the above plot, and shows a distribution of the error in distance from real to predicted locations. The distribution is skewed to the left, with 1 as the mode. And with the summary of the data, we can also see that