Elena Bonan – 27/01/2019

**REPORT**

Evaluate techniques for Wifi Locating

**GOAL**

The aim of our analysis is to see how machine learning can help to localise a person using wifi- fingerprints. In particular, we have considered a data set created in three muiltifloor buildings at the University of Jaume I in Spain and we have tested there the performances of our algorithms.

**AGENDA**

* Quick introduction about Wifi Fingerprint-Based Indoor Positioning
* Creation of the dataset
* Preprocessing
* Feature selection
* Modeling
* Identify the buildings
* Localisation in Building 1
* Localisation in Building 2
* Localisation in Building 3
* Conclusion
* Suggestions for future projects

**Wi-Fi Fingerprint-Based Indoor Positioning**

Wifi positioning system is a geolocalisation method used to locate the position of a device ables to recive wifi-signals. Knowing the characteristics of Wi-Fi hotspots and other wireless access points near the device and measuring the intensity of the signals recived, one tries to infer the position. However, unlike the GPS technology, it is difficult to obtain a good ‘approximation’. Indeed, the key challenge of this method, is overcoming the unpredictability of Wi-Fi signal propagation through indoor environments. Walls, objects, and even people cause the Wi-Fi signal strength to vary due to absorption, reflection, scattering, and diffraction. One way to approach the problem is to create a radio map: several empirical observations in the indoor enviroment considered. One can then use machine learning algorithms to create a model with these data. A new observation then will be located using this model.

**Creation of the dataset**



**Where**: It was created in three buildings of Universitat Jaume I with 4 or 5 floors and almost 110.000m2.

**When**: It was created in 2013 by means of more than 20 different users and 25 Android devices.

**Why:** The aim was to create a publicly available database for researchers

Figure 1: The three building examinated. From the left we have building 1, building 2 and building 3.

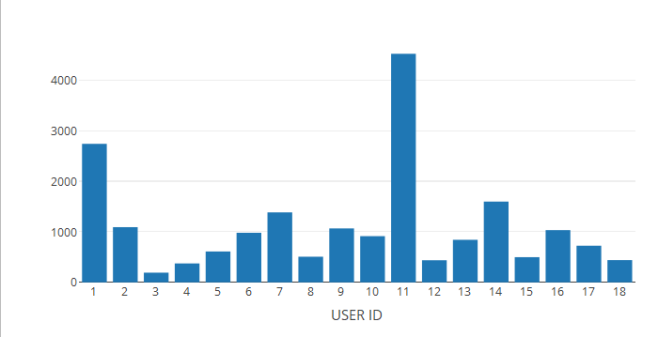
**Radio map:** It contains 19937 records. Regarding every observation, we have the detected RSSI levels of the 520 WAPS considered, the real word coordinates of the sample point, the building , the floor, the space ID (code to identify the rooms), relative position(if the sample was inside of the room or in front the entrance), the user, the phone and the time. The records are generated by 18 users. In every building were marked some positions where the users had to take captures. For every position each user had to take 10 captures. It was asked to cover each marked position by at least two users. No further indications were given. The users were free to captures the signals in their own way regarding time, position of the telephone with respect to the body, etc.

**Validation set:** It was created 4 months later then the training set. It has 1111 records. Respect to the training set, the fields that are missing are UserID, SPACEID and relative position. The records were generated by 14 users. They were free to capture the positions they wanted.

**Preprocessing**

Now we will explore the data set, looking at the distribution of the variables.

In the following histogram one can see the number of observations done by user ID.



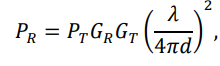
We see that the number of observations is not equally distributed between the users. It is important to keep in mind that the strength of the signals received depends also on the telephone used, the position of the phone respect to the user, the height of the user and so on. For this reason, the signals received in each position can depend on the person who took the capture and the way he took it.

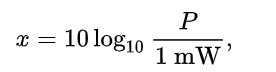
If we look at the phone ID we see that all the participants have a different model except for the users 1,9 and 19. They have the same model of telephone and android version. (Two of them shared the same device).

The first thing we did was to check if there were some WAPS whose signal was not received in any observation. We found 55 WAPS in the training and 153 WAPS in the validations.

Since we wanted to construct a model just using the training set and test it using just the validation, we could not consider these variables that where not ‘present’ in both sets. We don’t know the reason why we have this asymmetry. Maybe some WAPS were turned off during the captures of the training/validation or it was due to the particular positions of the people in the training and validation.

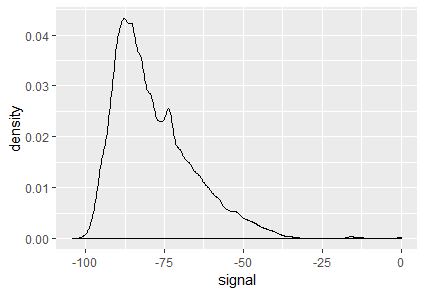
To have an idea about what it is the signal received, we have considered the following formulas. The first one explains the relationship between the power transmitted by the wifi device (Pt) and the power recieved by the telephone (Pr). Gr and Gt are two constants which depend on the devices, d is the distance between the phone and the router, λ is the length of the wave.



The RSSI is the following

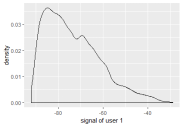
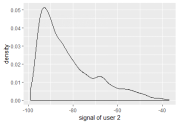
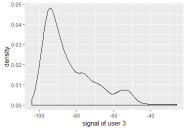
where P is the power received when the power transmitted is 1 mW. We see that the RSSI is not linear with the distance and that it is a negative number.

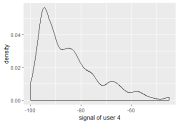
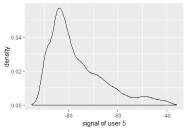
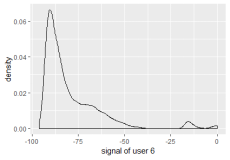
We wanted to have an idea about the usual signals detected by the participants. First, we do a density plot of the signals of the WAPS detected by all the participants. We omit the absence of signal which represent the 95% of the total signals.

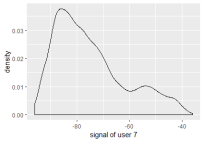
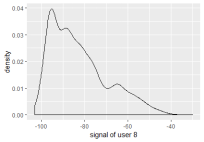
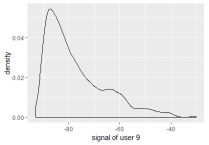


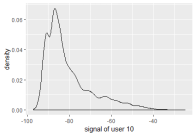
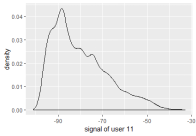
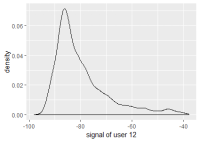
We see that the majority of the signals are around -90 dBm. It is strange that there are some signals greater then -30dBM. Indeed, usually one device can’t receive a signal of such intensity. It can be due to some anomalies or in the telephones used to take the records, or in the machine used to register the records, or some interference in the signals due to the environment etc.

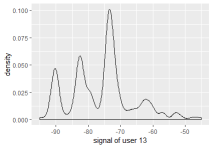
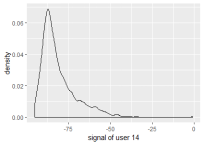
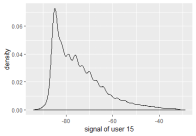
We plot the signals received by every user to check if there are some visible anomalies in the devices.

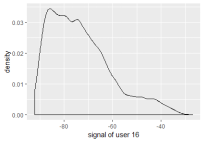
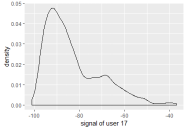
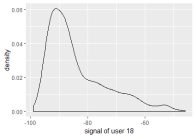
  

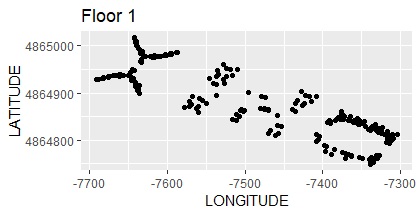
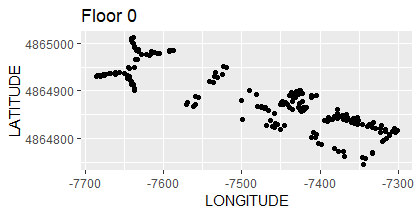
  

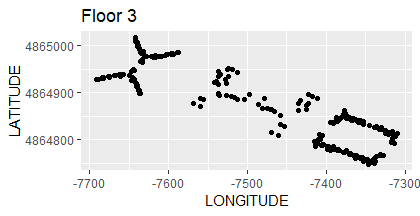
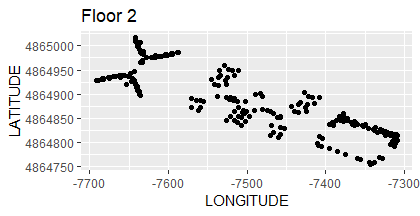
We notice that the user 6 received a signal of intensity greater than -30 more than all the other users. Probably there is a problem in that device. On possibility it is that it tends to amplify the signal when it is of high intensity. We decided to remove the observation of that telephone that have one signal greater than -30 dBm.

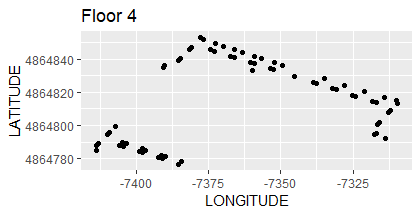
Furthermore, also the signals received by the user 13 present some anomalies. Since the density present three evident picks and the most frequent value is around -75 while the other users present just one high peak around -85.

It turned out that it is because there are a lot of duplicate observations for the user 13. That is, observations with the same value for every variable, also the time stamp. We decided then to eliminate the duplicate rows in the data set, solving the problem of the strange distribution of the signals for the user 13.

Now we look at the position where the signals were captured. We plot the position floor by floor.

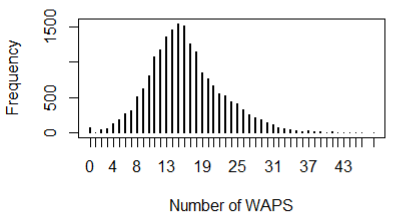




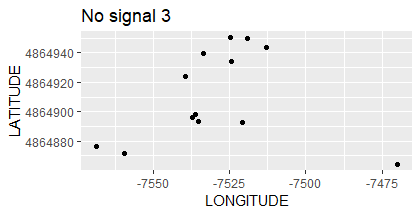
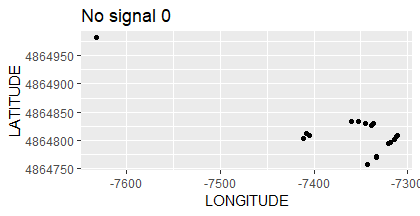


At a first look seems that the observations are distributed well in building 1. In building 2 we see that in the center of the building there are some zones that aren’t captured. Probably the users could not have access to some rooms or there are some architectural obstructions. In building 3 one part of the fourth floor is clearly missing.

Looking at the following graph, we can have an idea of the number of WAPS that are received for every observation.



The mean is 16.2, the max is 48 and min is 0. Having too much WAPS received could be a problem cause there could be too much noise. On the other hand few WAPS could not be enough to locate the device. There are also some observations with no signals. We are going to see where were made these observations to check if they provide some information regarding some specific place where there is no wifi connection or other problems. We have plotted the observations floor by floor. In the second and fourth floor there is just one observation without signal. More interesting are the observation in the first and third floor.



We just see that there is a concentration of observations in building 3 on the ground floor. However, there are a lot of other observation in the same position with signals. We conclude that these anomalies are due to a temporal problem.

Other characteristics to consider are the number of captures for position. There are places with just one observation and other with more than 50 captures. This can bring to a bias in the results since some positions weight more than other.

When we will analyse the building one by one we will consider again the following important factors:

* shape of the building
* distribution of the observation in the training set
* number of observations for every place in the training set
* users involved

We will then check how these factors impact the results.

**Feature selection**

We decided to divide the WAPS by building, i.e we try to identify where the WAPS were located. We eliminated the two WAPS which have signals in all the three buildings because the signal was low everywhere. For the other WAPS we used the following criteria. If the signal was received in only one building, say building 1, we ‘assigned it to that building’. If it was received by two buildings, say building 1 and 2. We assigned it to the building which have both the maximums signal received (at least -5 dBm more than the other building) and the maximum number of observations was in that building. With these criteria at the end we had 120 columns in building 1, 108 in building 2 and 76 in building 3.

**Modelling**

First we tried to locate the observations of the validation set in the right building using all the waps we selected before.

Then we constructed a model for every building. To create the training set we used the WAPS associated to that building and the observations made there. We predicted the floor, longitude and latitude in parallel.

For every prediction, we considered three classical machine learning algorithms: knn, random forest and support vector machine. We have tuned the parameters using the training set then we have compared the performances in the validation set. For the classification problems we have considered accuracy, kappa and confusion matrix . For the regression we looked at the RMSE, R squared, MAE and the plotted of the distribution of the errors. After have chosen a model, we considered longitude and latitude together. In a histogram we have the frequency of the distance between the real position in the validation instances and the position predicted. In a scatter plot for every instance of the validation, we have plot together the error in longitude and in latitude with sign.

**Identify the building**

The model that gave us better performance was knn with k = 1. These are the results.

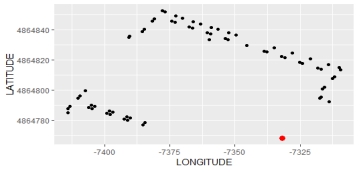
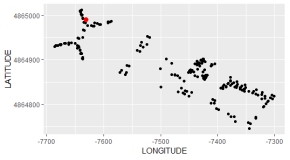
Real

|  |  |
| --- | --- |
| Accuracy | Kappa |
| 99.98 % | 99.98% |

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | 3 |
| 1 | 535 | 0 | 0 |
| 2 | 1 | 307 | 1 |
| 3 | 0 | 0 | 267 |

Predicted

We have that 1 person that was in building 1 in the ground floor was located in building 2 and 1 person that was in building 2 in the last floor was located in building one. Let us check where are these two persons (red points) respect to the positions of the points in the same floor in the training set (black point)



In the second case the person is in a place with no near observations in the training set. For the other case probably there was an isolate problem in that capture.

From now on we will not consider further these two observations of the validation set. Since the predicted building is wrong is not interesting to know the position predicted inside the wrong building.

**Building 1**

For this building we have 535 istances in the validation and 5245 observations in the training. After checked knn, random forest and svm we chosed the followings.

**FLOOR**

**Random forest**  Real

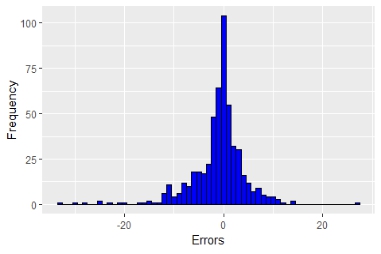
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 |
| 0 | 74 | 2 | 0 | 0 |
| 1 | 2 | 206 | 3 | 0 |
| 2 | 1 | 0 | 161 | 2 |
| 3 | 0 | 0 | 1 | 83 |

Predicted

|  |  |
| --- | --- |
| Accuracy | Kappa |
| 98% | 97% |

The results are quite good. Just one time we were wrong of two floors.

**LONGITUDE**

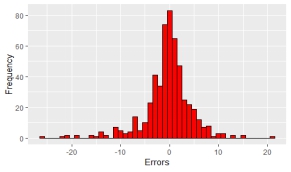
**Knn** k = 4

|  |  |  |
| --- | --- | --- |
| RMSE | R2 | MAE |
| 5.66 | 96% | 3.55 |

Even if the errors greater then 20 are more with the negative

sign, we are quite satisfied with the distribution.

The maximum error in module is 33 meters.

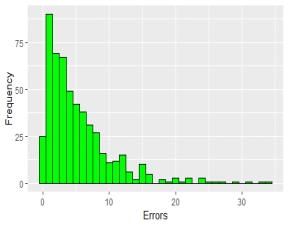
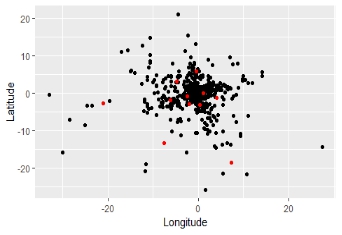
**LATITUDE**

**Knn** k = 4

|  |  |  |
| --- | --- | --- |
| RMSE | R2 | MAE |
| 4.93 | 98% | 3.22 |

Also, in this case the shape of the distribution is acceptable.

**Error in the position**

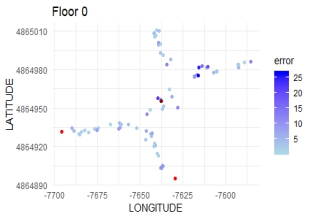
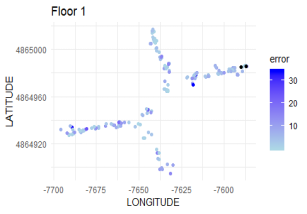
 

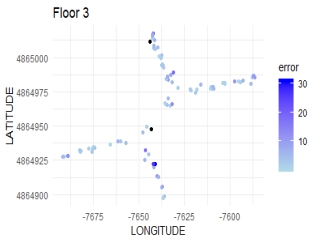
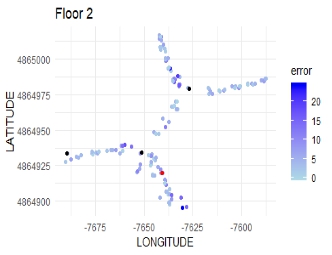
Right floor

Wrong floor

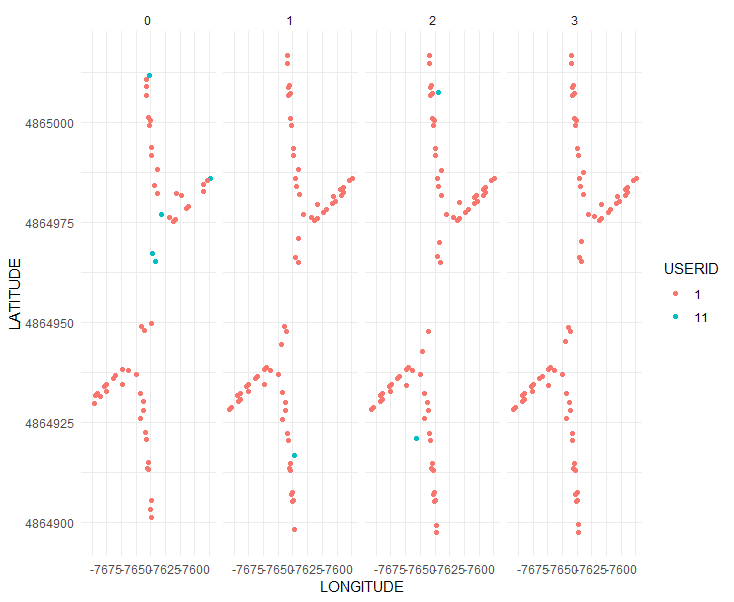
I have that the errors bigger than 10m represent the 14 % of the total. It means that with 86% of probability we managed to locate the person inside a circle of radius 10 meters.

Let us have a look at the positions of the errors in the following graph.

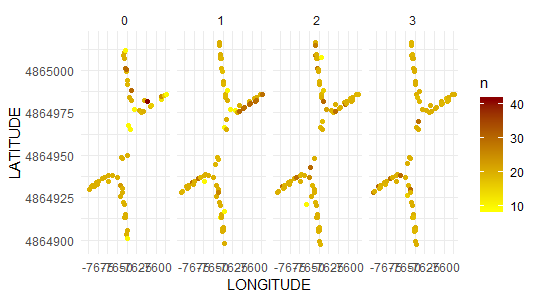
****

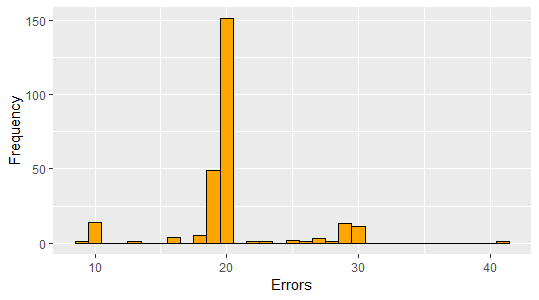


We see that in the extremities of the building sometimes we have picked the wrong floor. One possible explanation is that if we look at the points with similar WAPS, we obtain some points in the floor above or below because we have less observations near the right position. Other problems are in the center of the building probably due to the strange shape of the building.

We want to check if there are some relations between the errors and the user who made the trainig set. 

We see that almost ll the captures were made by the user 1. In the following graph are represented the number of captures per place.





**Captures per position**.

We see that there there is a position the ground floor with more then 40 captures. This can be a reason why we have some error in that zone.

**BUILDING 2**

For this building we have 307 instances in the validation and 4903 observations in the training.

Real

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 |
| 0 | 22 | 3 | 0 | 0 |
| 1 | 5 | 95 | 1 | 2 |
| 2 | 3 | 43 | 72 | 3 |
| 3 | 0 | 2 | 14 | 42 |

**FLOOR**

**Knn** k=1, kernel = optimal

Predicted

|  |  |
| --- | --- |
| Accuracy | Kappa |
| 75 % | 64% |

**Random forest** 100 trees, mtry = 10

Real

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 |
| 0 | 23 | 1 | 0 | 0 |
| 1 | 4 | 99 | 1 | 0 |
| 2 | 2 | 35 | 82 | 5 |
| 3 | 1 | 8 | 4 | 42 |

|  |  |
| --- | --- |
| Accuracy | Kappa |
| 80% | 71% |

Predicted

Real

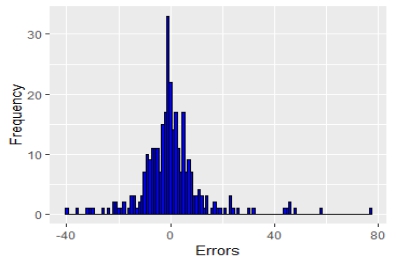
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 |
| 0 | 22 | 5 | 0 | 0 |
| 1 | 5 | 109 | 1 | 2 |
| 2 | 3 | 28 | 84 | 2 |
| 3 | 0 | 1 | 2 | 43 |

**SVM** kernel = Radial

|  |  |
| --- | --- |
| Accuracy | Kappa |
| 84% | 78% |

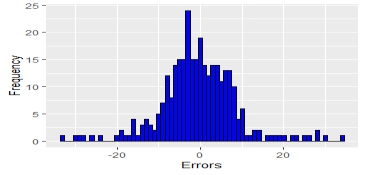
Predicted

We chose SVM because its performance was remarkable better than the other algorithms. We notice that in six cases we got an error bigger than two floors. We see clearly that there was a problem of classification between floor 1 and 2. More precisely 28 people that were in floor 1 were located in floor 2.

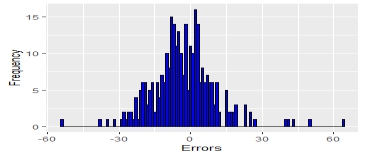
**LONGITUDE**

**Knn** k = 4, optimal

|  |  |  |
| --- | --- | --- |
| RMSE | R2 | MAE |
| 12.3 | 93 % | 7.58 |

**Random Forest,** trees = 100, mtry = 36

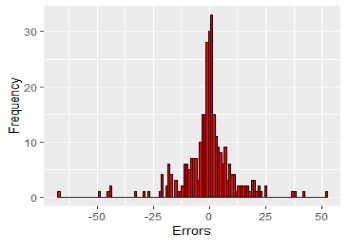
|  |  |  |
| --- | --- | --- |
| RMSE | R2 | MAE |
| 9 | 96% | 6.57 |

**SVM**

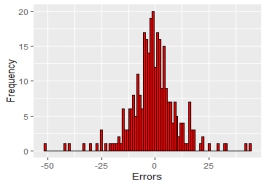
|  |  |  |
| --- | --- | --- |
| RMSE | R2 | MAE |
| 13.65 | 92% | 10 |

We chose the random forest. However, we are not completely satisfied with the distribution of the errors since it is not symmetrical. It is worth to notice that with knn there was a higher frequency of errors near 0 but there were few errors very big that decreased the overall performance of the algorithm.

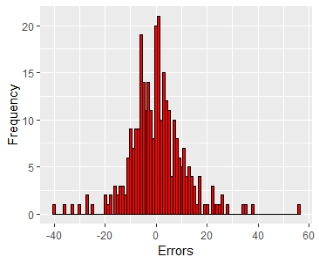
**LATITUDE**

**Knn**

|  |  |  |
| --- | --- | --- |
| RMSE | R2 | MAE |
| 12.03 | 88% | 7.39 |

**Random Forest**

|  |  |  |
| --- | --- | --- |
| RMSE | R2 | MAE |
| 11.06 | 90% | 7.71 |

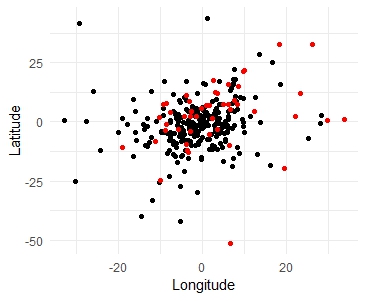
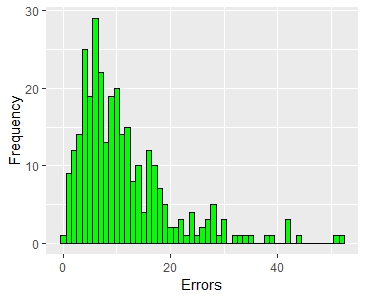


**SVM**

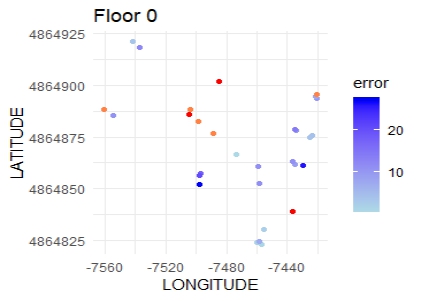
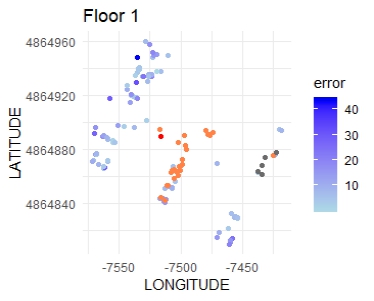
|  |  |  |
| --- | --- | --- |
| RMSE | R2 | MAE |
| 10.95 | 91% | 7.81 |

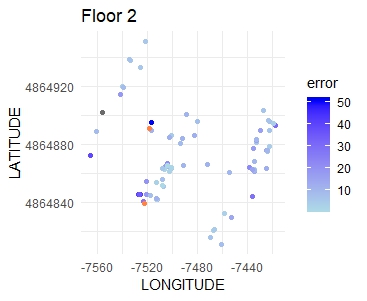
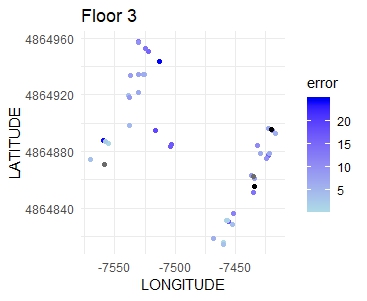
We chose the random forest even if the metric of errors were slightly worst than svm. The reason is that the distribution of the errors was more near to a normal distribution.

**Error in the position**

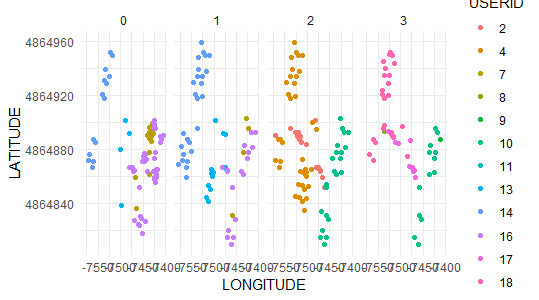


We see that in few cases we got an error greater than 40. Let us see where are located the errors.

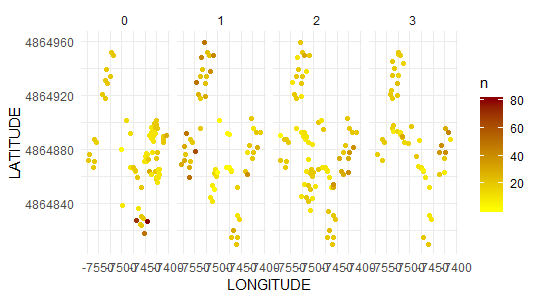
 

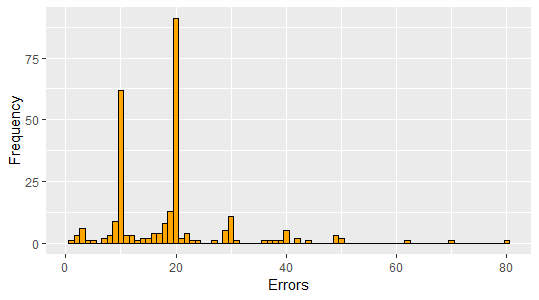
We have clearly a problem in the central region of the floor 1 since all the devices were located in the wrong floor. Still in the floor one there is an area which is predicted in the ground floor. Let us look at who took the captures

****

We see that respect to the building 1 we have more users involved. However, most of the time a zone is covered by just one participant. Then having more phone is no more an advantage.

Let us look at the number of observations for every reference point.

****

****

We see that with respect to the building 1 the situation is ‘worst’ indeed we have more variance.

**BUILDING 3**

For the third building we have 9076 observations in the training and 268 instances in the validation. Let us see the performance of the models.

Real

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 |
| 0 | 21 | 0 | 0 | 0 | 0 |
| 1 | 3 | 103 | 7 | 0 | 1 |
| 2 | 0 | 0 | 44 | 0 | 0 |
| 3 | 0 | 8 | 3 | 39 | 8 |
| 4 | 0 | 0 | 0 | 1 | 29 |

**FLOOR**

**Knn** k=1, kernel optimal

Predicted

|  |  |
| --- | --- |
| Accuracy | Kappa |
| 88 % | 84% |

**Random forest** 100 trees, mtry = 8

Real

|  |  |
| --- | --- |
| Accuracy | Kappa |
| 95% | 93% |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 |
| 0 | 21 | 0 | 0 | 0 | 0 |
| 1 | 3 | 110 | 1 | 0 | 1 |
| 2 | 0 | 1 | 51 | 0 | 0 |
| 3 | 0 | 0 | 2 | 39 | 5 |
| 4 | 0 | 0 | 0 | 1 | 32 |

Predicted

**SVM** kernel = Radial

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 |
| 0 | 22 | 0 | 0 | 0 | 0 |
| 1 | 2 | 109 | 6 | 0 | 3 |
| 2 | 0 | 1 | 46 | 0 | 0 |
| 3 | 0 | 1 | 1 | 39 | 3 |
| 4 | 0 | 0 | 1 | 1 | 32 |

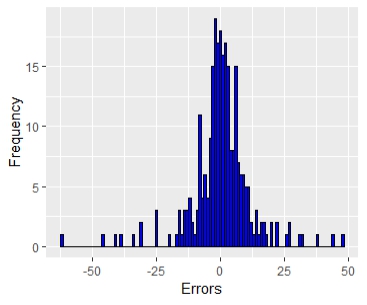
|  |  |
| --- | --- |
| Accuracy | Kappa |
| 93% | 90% |

Real

Predicted

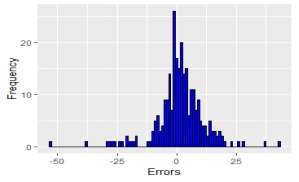
We chose the Random Forest. We have 1 error of three floors, probably the wrong position of the person was registered.

**LONGITUDE**

**KNN**

K = 5

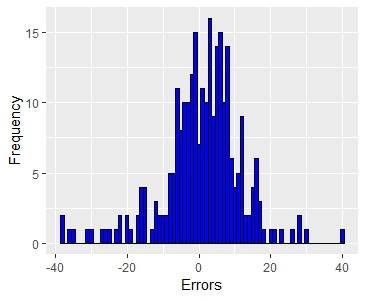
|  |  |  |
| --- | --- | --- |
| RMSE | R2 | MAE |
| 11.82 | 87% | 7.52 |

**Random forest**

Mtry = 25

|  |  |  |
| --- | --- | --- |
| RMSE | R2 | MAE |
| 10.23 | 90% | 6.86 |

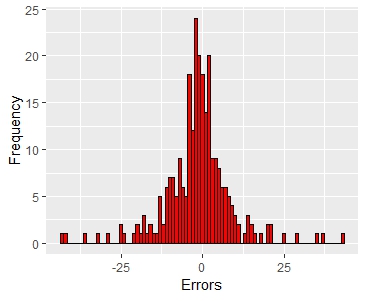
**SVM**



|  |  |  |
| --- | --- | --- |
| RMSE | R2 | MAE |
| 11.15 | 88% | 8.17 |

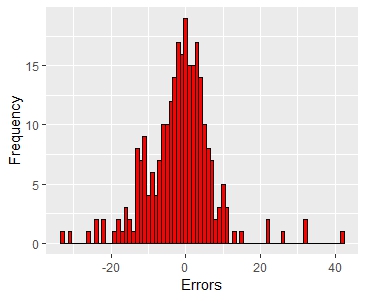
We chose the random forest.

**LATITUDE**



Knn k = 5

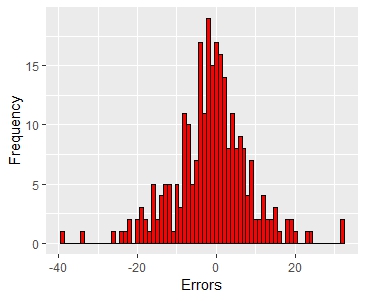
|  |  |  |
| --- | --- | --- |
| RMSE | R2 | MAE |
| 10.36 | 88% | 6.84 |



**Random forest** mtry = 8

|  |  |  |
| --- | --- | --- |
| RMSE | R2 | MAE |
| 9.06 | 90% | 6.34 |

**SVM** kernel = radial

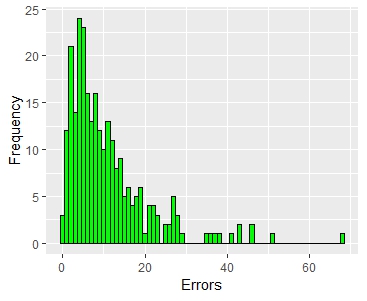
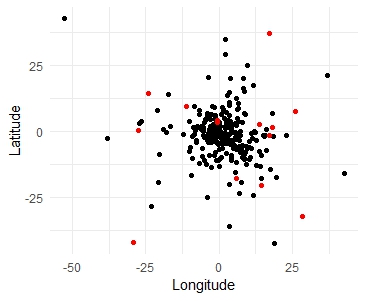


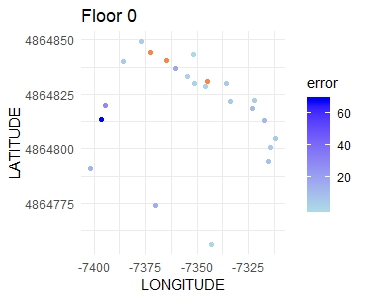
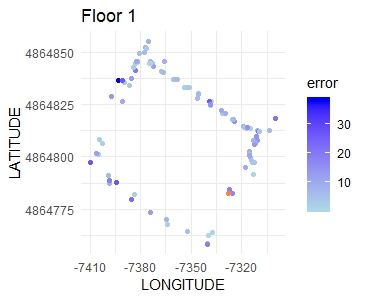
|  |  |  |
| --- | --- | --- |
| RMSE | R2 | MAE |
| 9.66 | 89% | 6.99 |

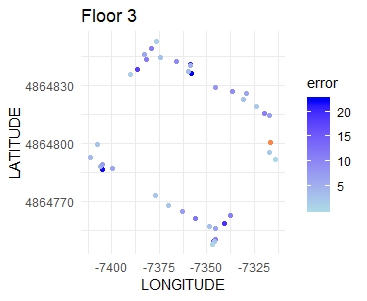
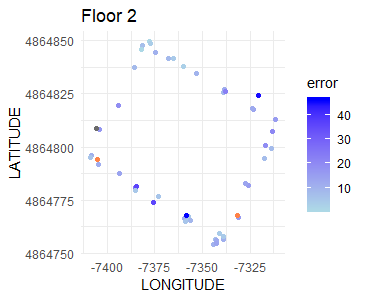
We chose knn.

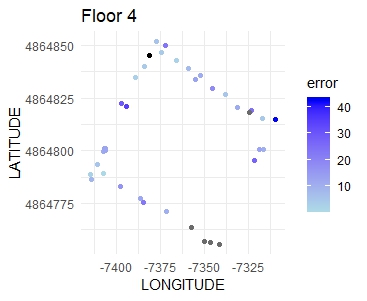
**Error in the position**

Let us see the frequency of the errors and the position.



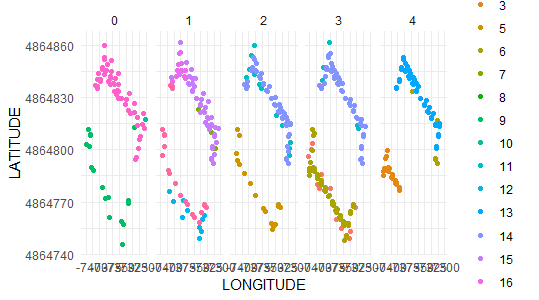
****

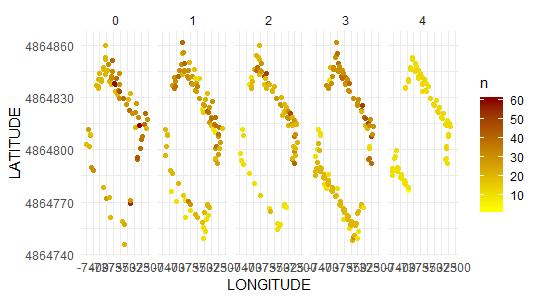
****

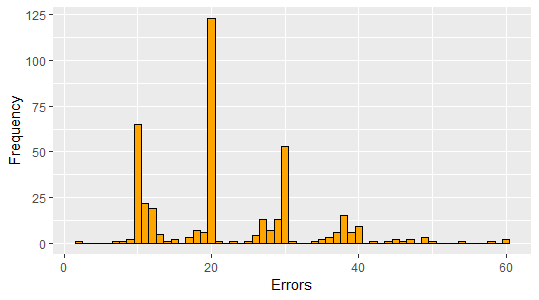
****

In the right-down-corner of the 4 floor we have that 4 observations were predicted in the third floor. The reason is that in that training there were no observation that zone. This tell us how important is to have observation in every zone.

The particular shape of the building, hole in the middle, can also represent a problem: the signal of two places can be similar even if they are distant (since the waves propagate better without walls etc.). Furthermore, we can predict a person to be in the middle that is outside the building.

****Let us see the users and the frequency of observations for the training set.

****

****

Just looking at these graphs we can not infer other interesting observations.

**CONCLUSION**

To sum up we have obtained the following results in the three buildings:

B1 B2 B3

Right floor: 98% 84% 95%

For right floor

MAE : 5.32 m 10.60 m 10.03 m

95 percentile: 14.90 m 27.84 m 26.73 m

90 percentile: 11.65 m 19.88 m 20.93 m

We believe that just the results for the first building are good enough for a concrete application of this locating system.

Since the validation set was created 4 months later than the training set we believe that there could have been sensitive changes in the WAPS. Some could have been removed or could have changed position.

Another important fact to consider is the shape of the building. Some positions which were more exposed seem to have been located worst.

We have found some clearly anomaly behaviour during the creation of the training set, like the strange signals of user 6 or the duplicates of user 13. But there could be other anomalies that we haven’t recognised.

**SUGGESTION FOR FUTURE PROJECTS**

For the generation of the radio map:

* Same number of observations per place.
* Same devices for every place. Possibly more than one.
* Declare the positions and the model of the wifi- routers

Try to apply more sophisticate machine learning algorithms:

* Clustered filter Knn
* Fuzzy C-means clustering
* Adaptive K-nearest point