Lara Cobler Moncunill
Ubiqum
27 January 2019

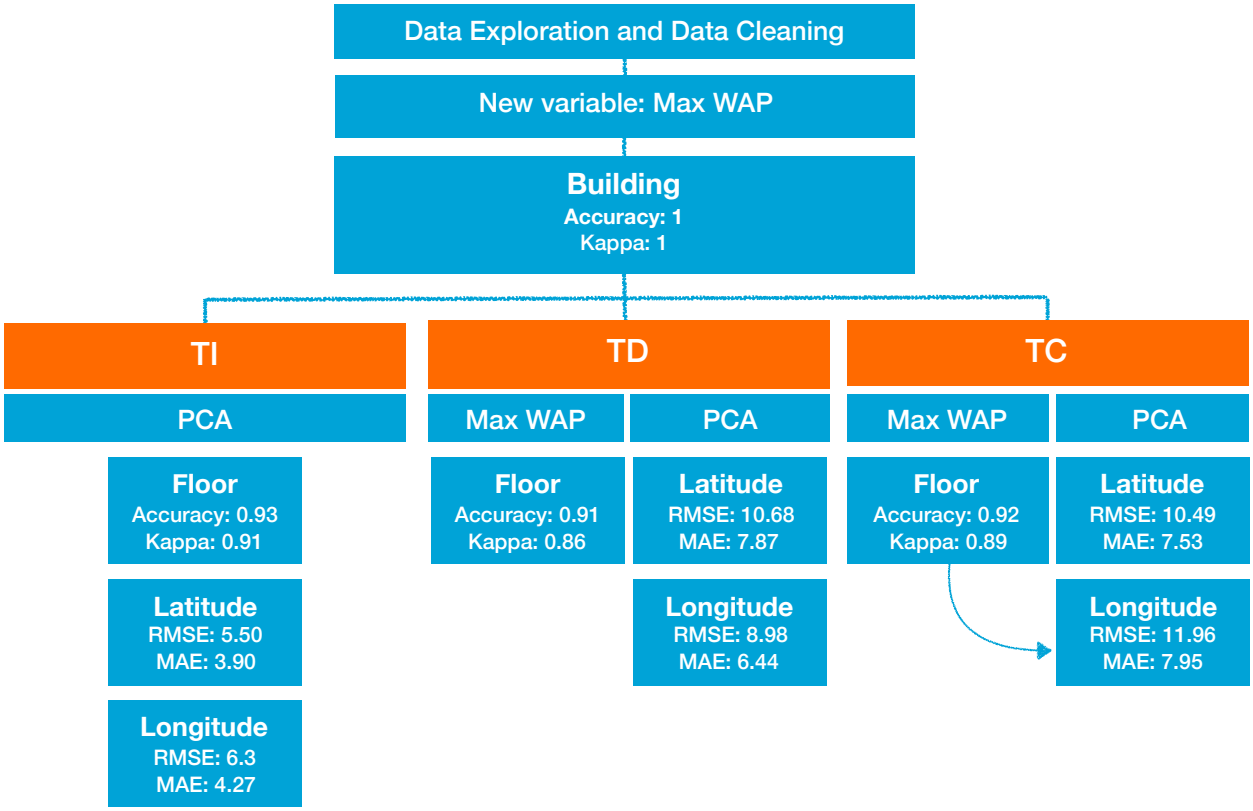
Indoors WIFI localisation

Summary

Goal

Use WIFI fingerprinting to determine a person's location in indoor spaces to develop a navigation system for interior spaces. In this cases develop a model to locate a person in three different buildings (TI, TD and TC) of Universitat Jaume I.

Analysis pipeline



Conclusions and Recommendations

The most influential factor in the performance of the model are quality RSSI signal and the phone used to get the fingerprint. To face this problems we suggest:

1. Add new WAPs to cover regions were RSSI signal is low.
2. Normalise RSSI value according to the phone
3. Mix data from training and validation to get more diverse training set to have more universal model.

Analysis

Goal

To develop a navigation system for indoors positioning using WIFI fingerprinting, a system that uses the signals from multiple WIFI hotspots within the building to determine the location analogously to how GPS uses satellite signals.

Dataset and data exploration

The database used is called UJIIndoorLoc (can be download at http://archive.ics.uci.edu/ml/datasets/UJIIndoorLoc) and covers three buildings of Universitat Jaume I with 4 or 5 floors and almost 110.000 m^2 (Figure 1).

Figure 1. Map of the three buildings covered for the dataset: TI-ESTCE (TI), TD-ESTCE (TD) and TC-ESTCE (TC).

The dataset consists of :

- Training set: 19937 training / reference records from 933 different locations (reference points).

- Validation set: 1111 validation / test records

Attributes description

- **RSSI levels of WAPs detected**.

A total of 520 WAPs (Wireless access points) that appear in the database (attributes 0 - 520) are represented by the RSSI (Received Signal Strength Indicator) value. The RSSI levels correspond to negative integer values in dBm, where -100dBm is equivalent to a weak signal and 0 dBm means extremely good signal. Generally, values higher than -30dBm are considered a bad lecture of the signal and values lower than -80 are considerer very low.

In the data set, each row corresponds to a fingerprint that records the RSSI received in a certain location. Both training and validation most fingerprints detects around 15 WAPs (Figure 2).

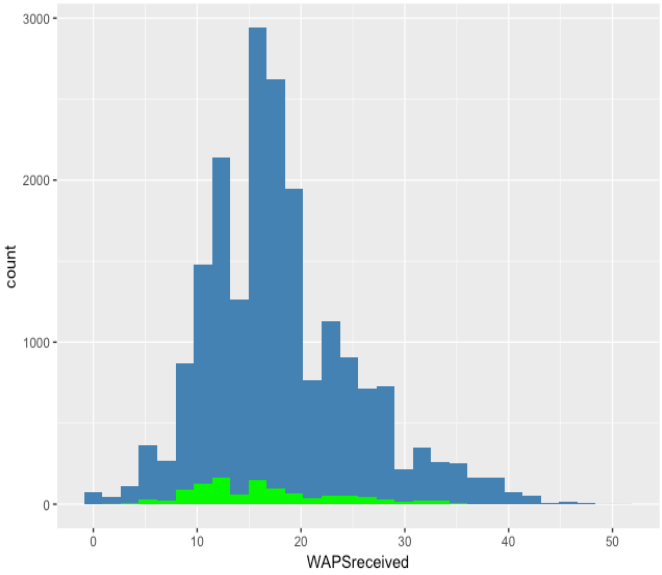


Figure 2: Histogram of WAPs received in each fingerprint.

- **Longitude and latitude coordinates**

Longitude and latitude coordinates in meters with UTM from WGS84 of each fingerprint.

- **Floor**

Floor were the fingerprint were recorded. The floor together with the longitude and latitude give the exact location of the fingerprint (Figure 3).

- **Building ID**

Identify the building, TI-ESTCE (TI), TD-ESTCE (TD) or TC-ESTCE (TC).

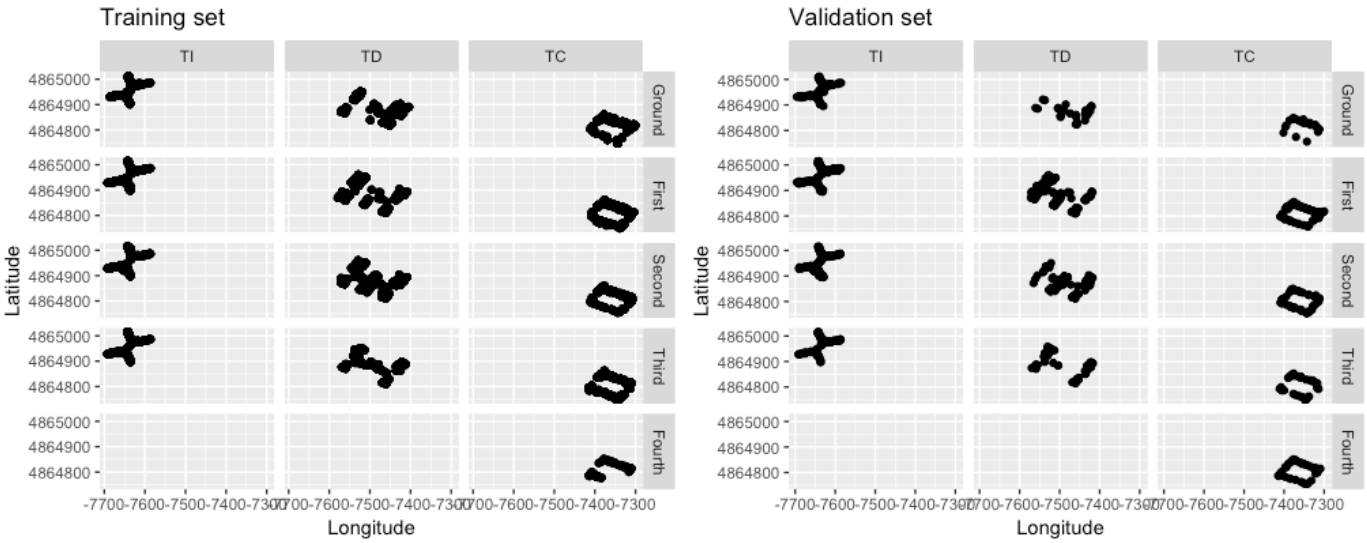


Figure 3: Fingerprints location by latitude (y axis), longitude (x axis), floor (vertical panels) and building (horizontal panels).

- **Space ID and Relative position**

Internal number to identify the space in the building (room number) and relative position respect to the space (inside / outside). Not provided in the validation.

- **User ID**

Eighteen different users for the training. Not recorded for the validation.

- **Phone ID**

Twenty-five different devices used considering the different Android version. Table 1 shows the user id that used each phone and user 0 corresponds to validation. From this we can see that 14 different phones were user only in the training, 9 only in the validation and only 2 in the training and validation. Also, for the training each phone only recorded fingerprints from specific regions, for example phone 13 (Figure 4) in the training for the middle building (TD) fingerprints were taken only in the left side of ground and first floor, whereas in the validation was used at random locations all over the building.

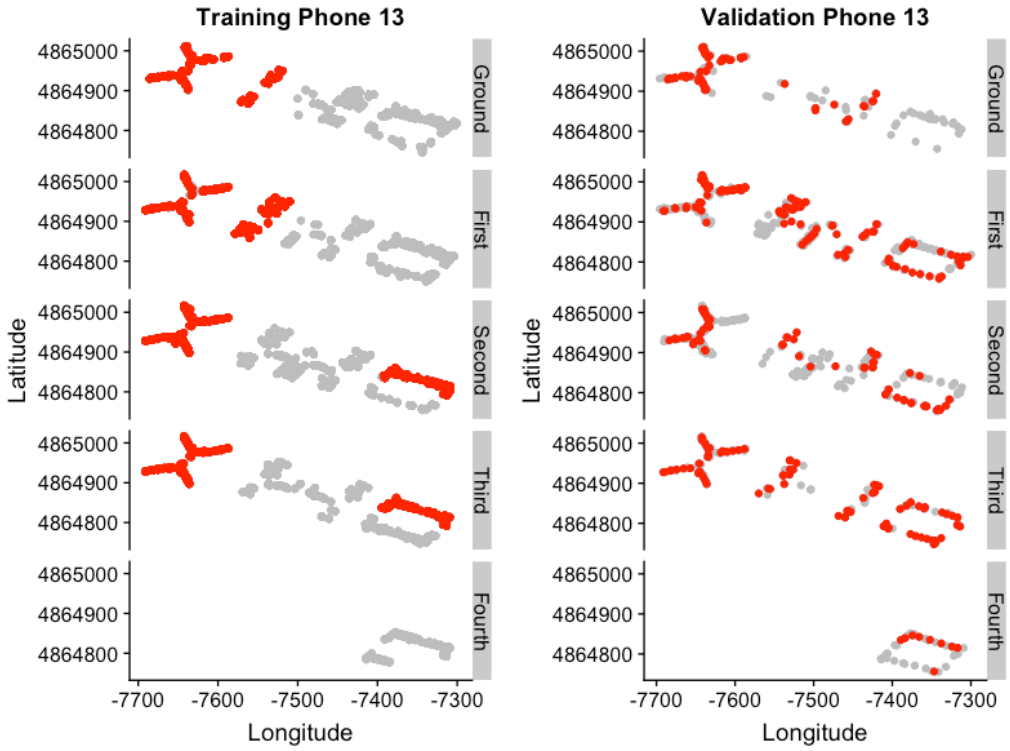


Figure 4: Fingerprints were the phone 13 was used (in red), for the training and validation set.

Table 1. Correspondence between Phone ID and real Android device and version. User who employed the device are also shown. Notice that user 0 corresponds to validation.

Phone ID	Android device	Android version	User
0	Celkon A27	4.0.4(6577)	0
1	GT-I8160	2.3.6	8
2	GT-I8160	4.1.2	0
3	GT-I9100	4.0.4	5
4	GT-I9300	4.1.2	0
5	GT-I9505	4.2.2	0
6	GT-S5360	2.3.6	7
7	GT-S6500	2.3.6	14
8	Galaxy Nexus	4.2.2	10
9	Galaxy Nexus	4.3	0
10	HTC Desire HD	2.3.5	18
11	HTC One	4.1.2	15
12	HTC One	4.2.2	0
13	HTC Wildfire S	2.3.5	0,11
14	LT22i	4.0.4	0,1,9,16
15	LT22i	4.1.2	0
16	LT26i	4.0.4	3
17	M1005D	4.0.4	13
18	MT11i	2.3.4	4
19	Nexus 4	4.2.2	6
20	Nexus 4	4.3	0
21	Nexus S	4.1.2	0
22	Orange Monte Carlo	2.3.5	17
23	Transformer TF101	4.0.3	2
24	bq Curie	4.1.1	12

- **Timestamp**

Corresponds to the UNIX time when the capture was taken. The data from the training set was obtained 6 different days from May 30th 2013 to June 20th 2013. May 30th, 31st, and June 4th, 10th and 12th collected data from TI building, and data from TD and TC building was only obtained June 20th.

Data from the validation set was obtained 9 different days from September 9th 2013 to October 8th 2013.

Data cleaning and feature engineering

The WAPs that were only detected in either training or validation were removed from the analysis. Those were 55 detected in the validation but not detected in the training, and 153 detected in the training set but out of order in the validation. Also WAPs that all signals were lower than -80 in the training (39 WAPs) were also excluded because lower than -80 is considered negative signal. The remaining 282 WAPs were used for the analysis.

The fingerprints that did not detected any WAP (76 from the training, none from the validation) and duplicates (637 from the training and 0 from the validation) were excluded from the analysis.

RSSI values higher than -30 are considered non-reliable, since -30 is the maximal signal reception. If we look at the fingerprints with some value higher than -30 most of them come from the phone 19 (84%). And 40% of all fingerprints obtained from phone 19 had some value higher than -30, this indicates that values obtained from this phone are not reliable and for this reason all fingerprints obtained from phone 19 were removed from the analysis (977 fingerprints).

In the training set, for the same location 10 different fingerprints were taken for one user with a certain phone, and at least another user with another phone recorded 10 more fingerprints at the same location. The 10 fingerprints taken at the same spot with the same phone were grouped as one, taking the median RSSI value for each WAP. At the end we have at least two fingerprints for each location obtained with different phones/users.

A new variable with the WAP that have the maximum signal for each fingerprint was created.

Insights from the data

One of the major pitfalls we may encounter is to locate the fingerprints with low signal, where all WAPs detected were lower than -80 (52 in the training and 13 in the validation). Also, to have a good 3D location we need at least 4 WAPs with RSSI higher -80, however there are 273 fingerprints from the training and 114 from the validation that does not accomplish this requirement. To be less strict we may only consider 3 WAPs with a signal higher than -80 to locate in the longitude and latitude, but still 218 fingerprints in the

training and 84 in the validation do not accomplish it. Figure 5 shows the location of the fingerprints that have less than 3 WAPs with RSSI higher than -80, less than 4 WAPs with RSSI higher than -80 and any WAPs with RSSI higher than -80, these locations are susceptible to have higher error in the location. It would be recommendable to add WIFI hotspots covering these areas to help with the localisation and also to have WIFI coverage in all locations.

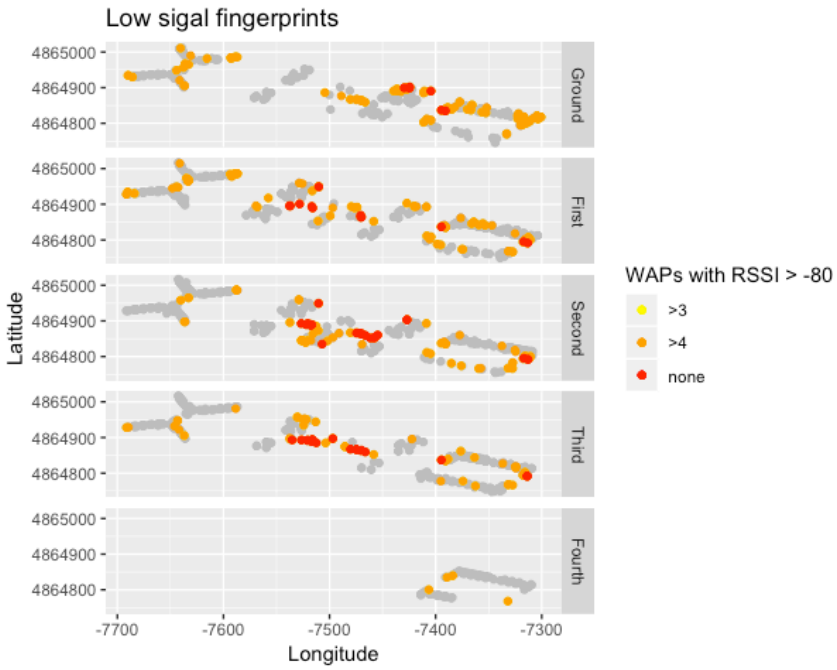


Figure 5: Locations with fingerprints with low RSSI signal.

At Figure 2 we can appreciate that there is a zone at the 4th floor of building TC that is not covered in the training set and there are some fingerprints in the validation set. Probably the model obtained from the training set will have a poor performance in predicting the location

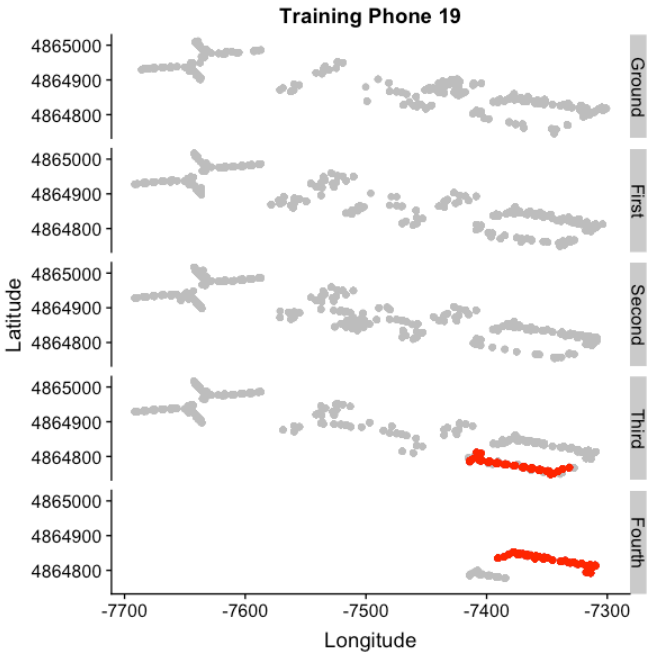


Figure 6: Fingerprints recorded from phone 19 that was excluded from the analysis.

of the fingerprints taken from that region. Also, the model would also have poor performance at 3rd and 4th floor of the TC building because the phone 19 excluded from the analysis was covering that region (Figure 6).

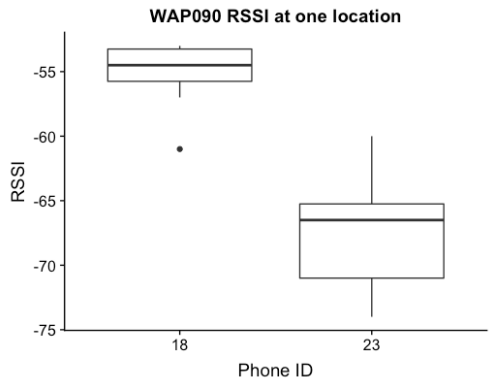


Figure 7: RSSI signal of WAP090 recorded with two different phones at the same location.

Other major influencing factors in to the model is the phone used to obtained the RSSI values. Figure 7 shows the RSSI signal obtained with different phones for the WAP090 at the same location, the RSSI value is clearly very different from the two different devices. This, together with the fact that the training different every device was covering specific regions of the buildings, and that different devices are used in the training and validation sets, clearly affects the performance of the model. It would have been better that each device used in the training would have record fingerprints equally distributed over the three buildings.

Modelling to predict location

Predict the building

Table 2 shows the error metrics from the models used to predict the building. A random forest with 50 trees was use to predict building using the RSSI value from all the WAPs (rf_building_all) and a random forest with 50 trees only using the variable that contains the maximal WAPs from each fingerprint (rf_building_max). The last one predicted all buildings from the validation set correctly, so it was selected to predict the building.

Because the building was predicted without errors, the dataset were split into the different buildings to have specific models to predict the floor, longitude and latitude for each buildings. The reasoning behind that is that different buildings have different shape and other factors that may influence in the RSSI reception making each building different.

Table 2: Models and error metrics to predict the building

Model	Accuracy	Kappa	Notes
rf_building_all	99.99	99.86	1 error
rf_building_max	**1**	**1**	**0 errors**

Principal Component Analysis (PCA) for each building

To reduce the dimensionality of the data to predict the floor, longitude and latitude of each building, the PCA of each building was performed and the first principal components (PC) that explain 90% of variance of each building were further used for the floor, longitude and latitude modelling. For the TI the firsts 18 PC were chosen, and the firsts 19 PC for the TD and TC.

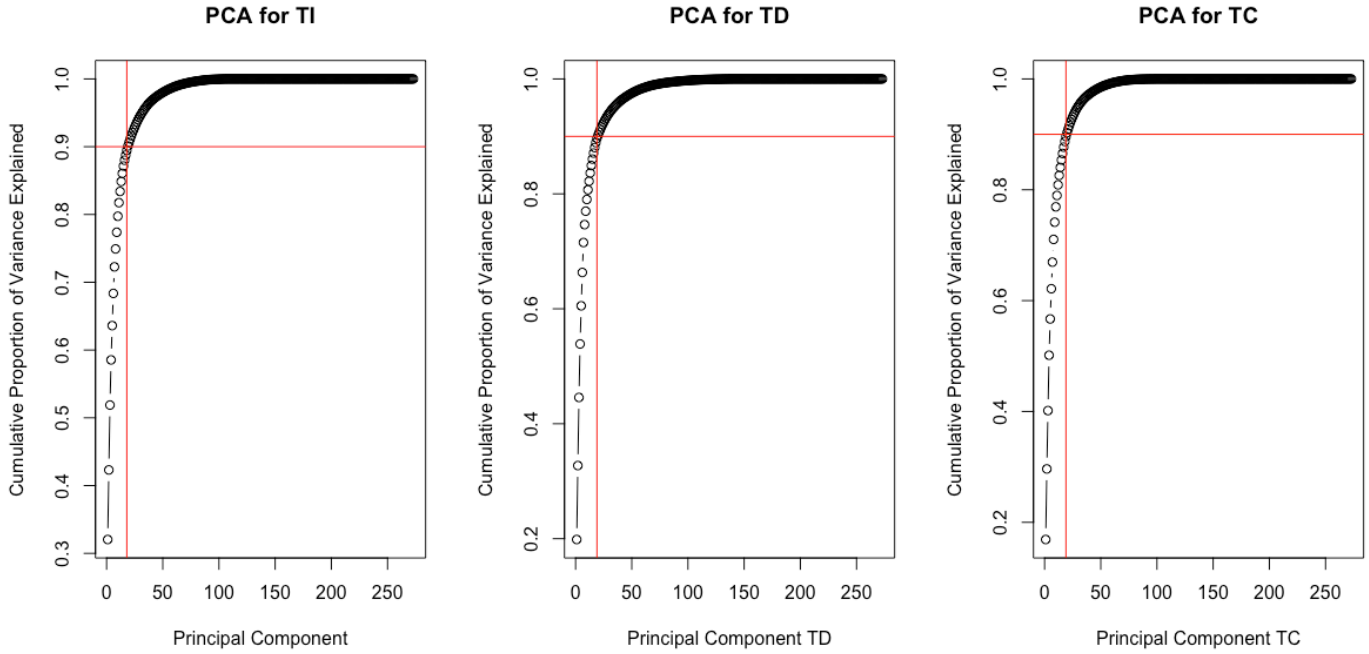


Figure 8: PCA for each building TI, TD and TC. The firsts principal components that explains 90% of the variance were used for further modelling (red).

Predict the floor for each building

To predict the floor of each building random forest with 50 trees and KNN were trained using either maximum WAP detected, all WAPs and the selected PCs. Always the random forest performed better in the training that the KNN, for this reason only the random forest models were used to test the performance in the validation set. Table 3 shows the models and their performance metrics in predicting the floor for each building.

To predict the floor in the TI building, although the model using all WAPs (floor_rf_ti_all) had higher accuracy and kappa (95.42 and 93.52, respectively) it had 4 errors of 2 floor distance (3 from the ground were predicted at the second floor, and 1 from the second floor was predicted at the ground floor). For this reason, the selected model was random forest using PCA (floor_rf_ti_pca, 93.5% accuracy and 91.1% kappa), that only situated 1 fingerprint from the ground floor at the second floor, the other errors were only one floor distant.

To predict the floor for the TD building the random forest using the WAP with maximum signal was selected (floor_rf_td_max, accuracy 90.88% and kappa 86.19%), only one error was 2-floor distant, from the ground floor was predicted at the second. The models using all WAPs and PCA presented most errors from the first floor.

To predict the floor for the TC building, the random forest using the maximum WAP for each fingerprint was selected (floor_rf_tc_max, 91.98% accuracy and 89.29% kappa), of 21 errors only 2 were 2-floor difference (from the first floor to the third floor). The models using all WAPs and PCA made most of errors with the fingerprints from the fourth floor, that had a region that was not covered in the training and the other region was cover with the phone 19, which was excluded (discussed later).

Table 3: Models and error metrics to predict the floor.

Building	Model	Accuracy	Kappa	Notes
TI	floor_rf_ti_max	0.8645	0.8092	4 at 2nd predicted at ground
	floor_rf_ti_all	0.9542	0.9352	3 in the ground predicted at the 3rd, 1 second at the ground
	floor_rf_ti_pca	**0.937**	**0.911**	**1 in the ground predicted at 2nd.**
TD	**floor_rf_td_max**	**0.9088**	**0.8619**	**1 in the ground predicted at 2nd.**
	floor_rf_td_all	0.772	0.6759	Lots errors from the first
	floor_rf_td_pca	0.7655	0.6644	Lots errors from the first
TC	**floor_rf_tc_max**	**0.9198**	**0.8929**	**2 from first predicted third**
	floor_gbm_tc_max	0.76	0.70	8-17 from 2nd, 3rd, 4th predicted at ground
	floor_rf_tc_all	0.8359	0.7778	26+1 errors from the 4th.
	floor_rf_tc_pca	0.8779	0.8335	All errors from the 4th floor (13+1+1)

Predict longitude and latitude for each building

To predict the longitude and latitude of each building random forest with 50 trees and KNN were trained using either selected PCs, all WAPs and WAPs located in each building. In all three buildings models, KNN or random forest, using selected PCs performed better in the training set than using all WAPs and specific WAPs of the building. Then, these models were optimised and SVM radial and gbm were also trained using the selected PCs for each building.

Table 4 shows the error metrics of the tested models, models that presented lower errors and as less as possible predictions out of the building were selected. In this case, both longitude and latitude the best model was KNN (for longitude: lon_knn_ti_pca; for latitude:

lat_knn_ti_pca). When combined the predicted coordinates for longitude and latitude the mean error is 6.5m.

Table 4: Models and error metrics to predict longitude and latitude for TI building

	Model	RMSE	squared-R	MAE	Notes
Longitude	**lon_knn_ti_pca**	**6.33**	**0.94**	**4.27**	**Good overlay, maybe a couple out of the building, Good distribution of errors.**
	lon_svmr2_ti_pca	8.63	0.90	6.10	Some out, no peaks at the ends, best regarding distribution of errors.
	lon_rf_ti_pca	7.51	0.92	5.10	Good overlay but some places out
	lon_gbm_ti_pca	7.62	0.92	5.11	Good overlay, some places out
Latitude	**lat_knn_ti_pca**	**5.50**	**0.97**	**3.90**	**Some out, but others overlay well**
	lat_svmr2_ti_pca	6.99	0.95	5.12	Lower values out of range
	lat_rf_ti_pca	6.54	0.96	4.32	good overlay in general, but some places out of the building
	lat_gbm_ti_pca	6.25	0.96	4.30	Good overlay, but few out of place

For the TD building the model selected to predict the longitude (lon_rf_td_pca) had the best error metrics and most of predicted locations seem to be in the building. For the latitude although KNN model had lowest MAE, some predicted locations were out of the building and had large error; for this reason, the chosen model was gbm (lat_gbm2_td_pca, RMSE 10.68, squared-R 0.91, MAE 7.87). When combined the predicted coordinates for longitude and latitude the mean error is 11.10m.

Table 5: Models and error metrics to predict longitude and latitude for TD building

	Model	RMSE	squared-R	MAE	Notes
Longitude	lon_knn_td_pca	10.20	0.95	6.93	Good overlay, maybe a couple out of the building
	lon_svmr2_td_pca	12.03	0.93	9.18	Some out
	lon_rf_td_pca	**8.98**	**0.96**	**6.44**	**Good overlay, some places out**
	lon_gbm_td_pca	9.20	0.96	6.57	Good overlay, some places out
Latitude	lat_knn2_td_pca	11.12	0.90	7.29	A lot out of the building
	lat_svmr2_td_pca	10.95	0.91	8.18	Some out of the building
	lat_rf2_td_pca	10.97	0.91	8.01	Good overlay, some places out of the building
	lat_gbm2_td_pca	**10.68**	**0.91**	**7.87**	**Good overlay, some places out of the building**

To predict the longitude of the TC building KNN using the PCA (lon_knn2_tc_pca) performed best in floors ground to third but at the fourth floor located most fingerprints in the middle of the building. When adding the predicted floor to the PCA for the prediction (lon_knn3_tc_pca), even though there are 8 fingerprints from the forth floor classified as third floor, the performance of the model improved, specially for the fourth floor (Table 6). To predict the latitude the model with best performance was random forest (lat_rf_tc_pca, RMSE 10.49, squared-R 0.88 and MAE 7.53). When combined the predictions for longitude and latitude for the TC building the MAE was 12.20m.

Table 6: Models and error metrics to predict longitude and latitude for TC building

	Model	RMSE	squared-R	MAE	Notes
Longitude	lon_knn2_tc_pca	13.68	0.81	8.84	Lot's error in the 4th floor
	lon_knn2_tc_pca For 0-3 floors	13.10	0.82	8.44	Still some in the middle
	lon_knn3_tc_pca	**11.96**	**0.86**	**7.95**	**PCA + Floor. Best 4th floor**
	lon_svmr2_tc_pca	13.75	0.81	10.39	Errors at the 2nd and 4th floor, wide error distribution
	lon_rf_tc_pca	13.19	0.83	9.33	Errors in the middle. Not bad 4th floor
	lon_rf_tc_pca for 4th floor	14.21	0.85	10.93	Right side all in the middle
	lon_gbm_tc_pca	13.23	0.82	10.01	Errors 2nd and 1st floor, wide error distribution
	lat_knn2_tc_pca	13.10	0.81	7.77	Errors at 4th and 2nd
	lat_knn3_tc_pca	12.71	0.82	7.75	+floor, still bad 4th and 2nd
	lat_svm2_tc_pca	12.67	0.81	9.07	Errors 2nd and 4th
	lat_svm3_tc_pca	12.59	0.81	9.17	+floor, still bad
	lat_rf_tc_pca	**10.49**	**0.88**	**7.53**	**Still bad 4th but much better than the others**
	lat_gbm_tc_pca	10.66	0.87	7.76	Similar to RF

Error location of the predictions

Figure 9 shows the predictions for floor, longitude and latitude of the building TI. At the top panel we can see the errors in the floor prediction, in the second floor, where most of the WAPs are located (table 7), it has less misplaced fingerprints. Also most of the errors

fingerprints recorded with phone 20 (10 of 33) and phone 0 (9 of 33), neither of them used for the training set (for this building only phone 13 and phone 14 were use for the training). Moreover, the phone 20 is a Nexus 4, the same model that was excluded from training set for unreliable RSSI values. The middle panel of figure 9 shows the real longitude and latitude coordinates of the fingerprint (base of the arrow) and the predicted location, all predicted seem to be inside the same building. Like in the floor the second floor has less position error than the other floors. The first and third floor in general have small errors but have a couple fingerprints in each floor with error around 30m. This fingerprints also were misclassified with the floor (bottom panel of figure 9), and 3 of this 4 fingerprints had less than 4 WAPs were the RSSI signal was higher than -80, necessary for the localisation, we may consider this low quality fingerprints. The other fingerprint with elevated error and floor misclassified was recorded with the phone 20. Also, the first floor has a zone at the bottom left with some errors in location, that seems to be taken also with the phone 20. In conclusion, the most influencing factor in the prediction of the TI building are the quality of the fingerprints (need more than 4 WAPs with RSSI higher than -80) and the phone used, the phone 20 seem not to be accurate.

Table 7: Number of WAPs per floor.

	TI	TD	TC
Ground	15	44	12
First	28	23	19
Second	32	21	14
Third	22	18	17
Fourth	-	-	8

Figure 10 shows the predictions for floor, longitude and latitude for TD building. At the top panel we can appreciate that most errors are at the ground floor predicted to the first and at the second floor predicted also to the first. This suggest that the WAPs at first floor have higher RSSI and the WAPs from ground and second floor have a weaker RSSI. For example 3 fingerprints from the ground floor predicted at first, and 2 from the second predicted at the first the WAP with maximum signal that detected was WAP037, located at the first floor. On the other hand, most of the errors for the floor are made by phone 20 (6 of 28), phone 5 (6 of 28), and phone 13 (6 of 28). Like at TI building, phone 20 and phone 5 were not used to get lectures for the training, and phone 13 was used for the training only left part of TD building of ground and first floor (Figure 4), whereas most of error made in the validation for this phone are out of this region. Regarding the predictions for longitude and latitude (middle panel of figure 10), all positions seem to be predicted inside the building. Seem to be a region at the end of middle wing of the second floor were all fingerprints have an error about 30m. All these fingerprints had less than 4 WAPs with RSSI signal higher than -80 (lower panel from figure 10), suggesting that this zone should be better covered with additional WAPs. There are also isolated fingerprints at the second and third floor with and error for position about 40 meters and, this fingerprints were recorded with the phone 20. In conclusion, the main factors affecting the performance of location for our model in the TD

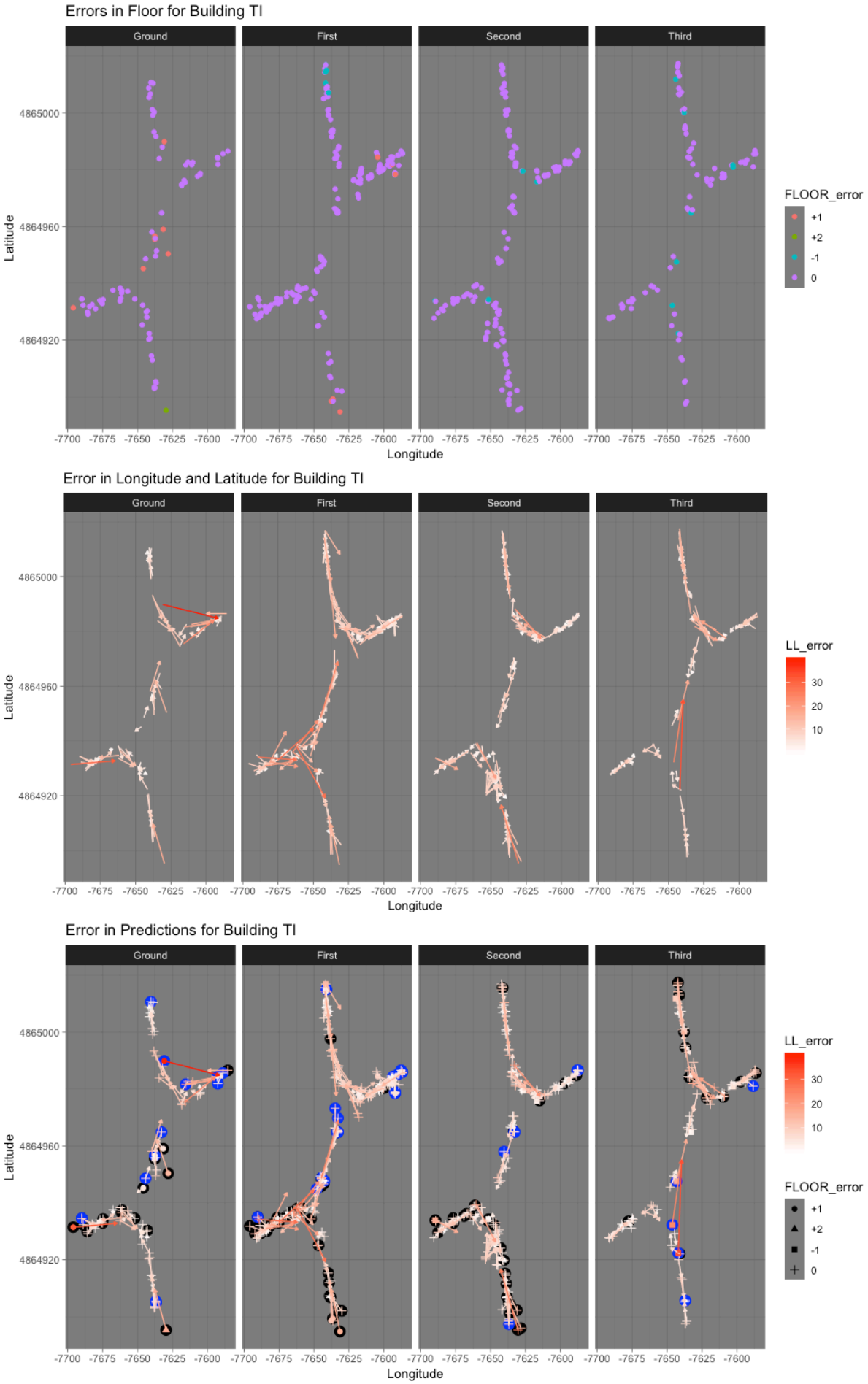


Figure 9: Top: Errors in the floor prediction. Middle: Base of the arrows are the real positions and the tip the predicted position. Bottom: Predictions for the position (same as middle panel), shape of the base indicates error in the floor prediction, black dots indicate fingerprints taken with phone 20 and blue dots indicate fingerprints with less than 4 WAPs with RSSI higher than -80.

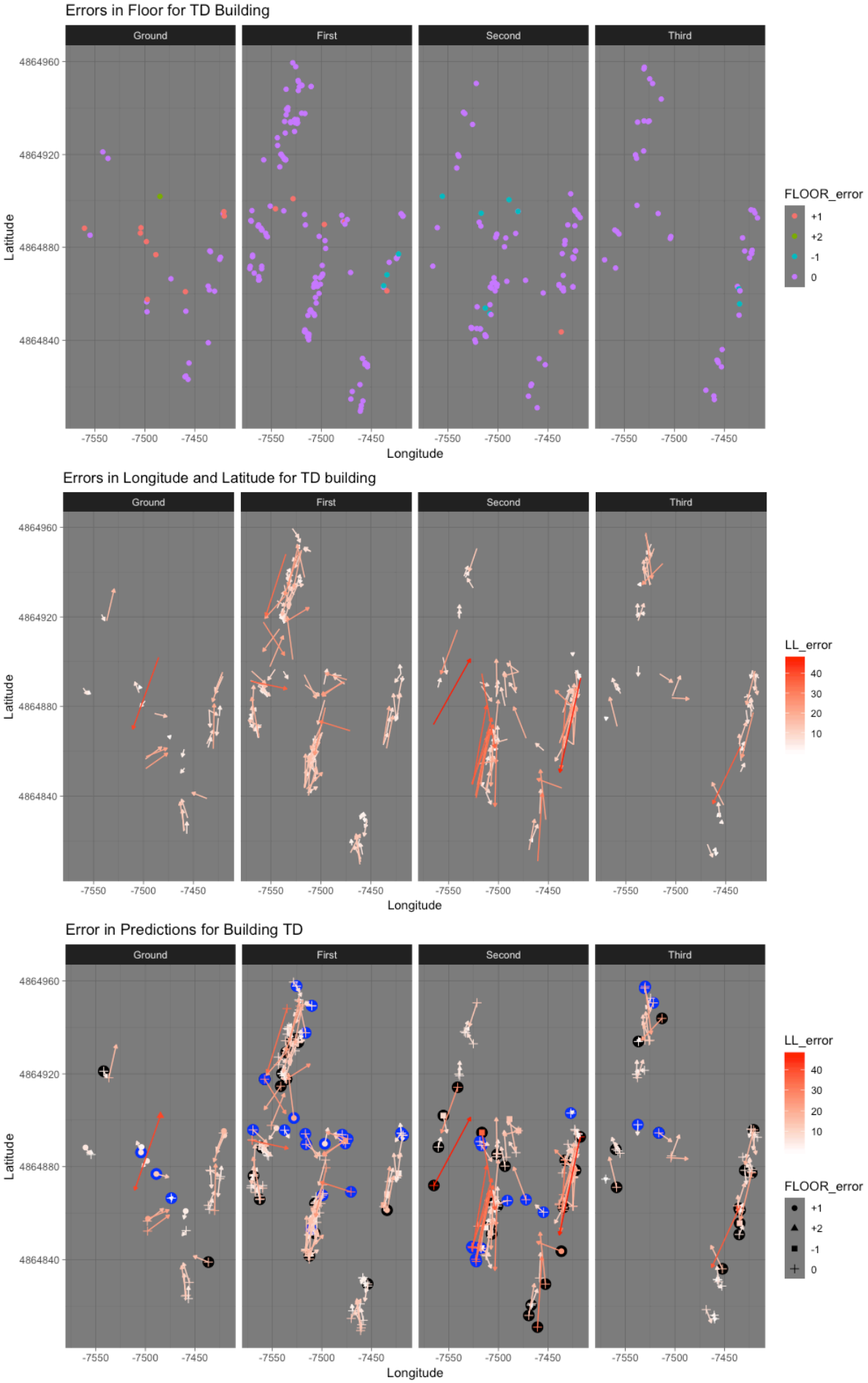


Figure 10: Top: Errors in the floor prediction. Middle: Base of the arrows are the real positions and the tip the predicted position. Bottom: Predictions for the position (same as middle panel), shape of the base indicates error in the floor prediction, black dots indicate fingerprints taken with phone 20 and blue dots indicate fingerprints with less than 4 WAPs with RSSI higher than -80.

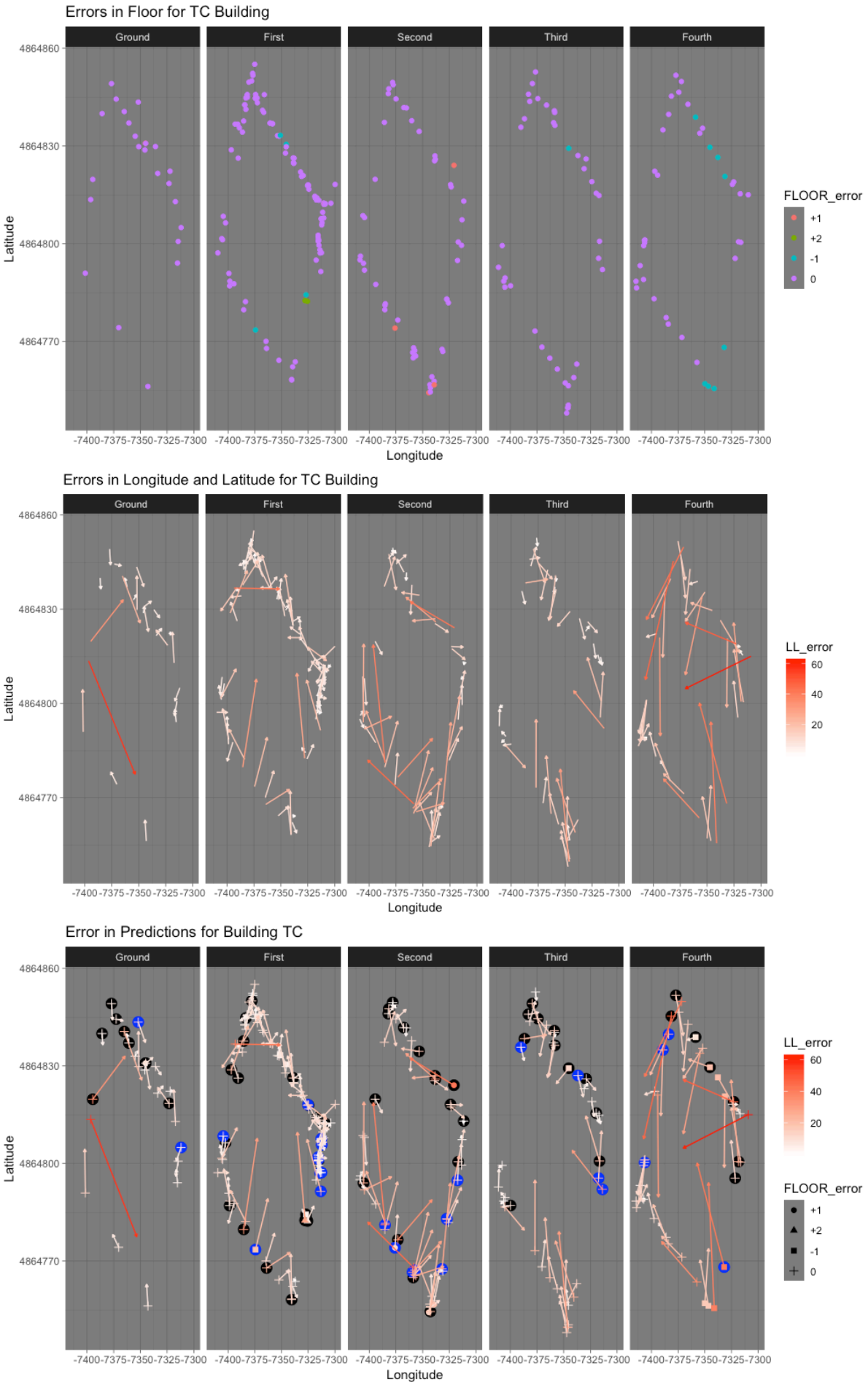


Figure 11: Top: Errors in the floor prediction. Middle: Base of the arrows are the real positions and the tip the predicted position. Bottom: Predictions for the position (same as middle panel), shape of the base indicates error in the floor prediction, black dots indicate fingerprints taken with phone 20 and blue dots indicate fingerprints with less than 4 WAPs with RSSI higher than -80.

building are the same as in the TI building, the quality of fingerprints and phone use to get them.

Figure 11 shows the errors in the location for building TC. The top panel shows the errors for floor location, the first and fourth floor are the ones that accumulate more misclassifications. And, again, most errors are fingerprints recorded with phone 20 (8 of 21), which we have already defined as unreliable. Other errors come from fingerprints recorded with phone 13 and phone 14 (6 and 4, respectively), although they were used in the training, the location of the positions that made the error were in different regions where this phones were used for the training. Moreover, most of the errors are at the fourth floor where there was a region not covered in the training and, also was covered with phone 19, which was excluded. Regarding the error in the position (longitude and latitude), most errors are at second and fourth floor (middle and lower panel of figure 11), which are regions not covered with the training (fourth floor), or regions with low RSSI signals fingerprints (figure 5).

Conclusions and Recommendations

The most two influencing factors on the performance of the model are the quality of the fingerprints and the phone used. Figure 5 shows the locations of low RSSI signal fingerprints, to add new WAPs covering these regions would certainly decrease the error in the location. Also if the RSSI is low for location in these regions, it is low for internet connectivity, Adding new WAPs to these locations would help with the localisation and the internet connectivity.

The other factor that influences is the phone used to record the fingerprints, specially phone 19 and 20 (Nexus 4), that are considered non relievable, for this reason phone 19 was removed from the training but then most of errors in the validation were with fingerprints obtained with phone 20. Other problems with the phones are that for the same location two different phones have different RSSI values, to improve the performance the RSSI values should be normalised according to the phone used.

For the training, every user with a certain phone was assigned to a specific region, to have a universal model that makes minimum error in any location with any phone, the phones used in the training should have recorded fingerprints more distributed all over the building to cover. In this same direction, for the training data for TD and TC buildings was obtained the same day, which also may influence with the RSSI recorded. To solve this problems of distribution in the training, we may mix fingerprints from training and validation to get more universal model regarding the phones used, the phone distribution and also different days, where the WAPs distribution may change. However, then we need a new validation set to prove that the performance in different conditions (time, phones, etc.) is better than the current model.