INDOOR LOCATIONING WIFI FINGERPRINT

JANUARY 27, 2020

Matias Barra

CONTENTS

Summary

Many real world applications need to know the localization of a user in the world to provide their services. Automatic user localization consists of estimating the position of the user (latitude, longitude and altitude) by using an electronic device. While Outdoor localization problem can be solved very accurately thanks to the inclusion of GPS, indoor localization is still an open problem mainly due to the loss of GPS signal in indoor environments.

Data Set

The UJIIndoorLoc database covers three buildings of Universitat Jaume I, and it records the signals of wireless access points received by cellphones according to their location (fingerprint).

The main characteristics of the dataset are:

Covers a surface of 108.703 m^2 including three (3) buildings with four (4) or five (5) floors.

The number of unique locations (reference points) in the database is 933.
21.049 sampled points have been captured: 19.938 for training/learning and 1.111 for validation/testing. Dataset independence has been assured by sampling the Validation set four (4) months after the training set. The number of different wireless access points (WAPs) appearing in the database is 520.
Data was collected by more than twenty (20) with twenty-five (25) different models of mobile devices.

The attributes provided in the dataset are:

(001-520) - RSSI Levels: These values represent the Received Signal Strength Indication (RSSI) level for each of the WAPs detected in a location. Each fingerprint is depicted as a 520-element vector.

The RSSI levels correspond to negative integer values measured in dBm, where −100dBm is equivalent to a very weak signal, whereas 0dBM means that the detected WAP has an extremely good signal.

It is important to note that not all the WAPs are detected in each scan (fingerprint), so an artificial value of +100dBm is used by default in those WAPs that have not been detected by the device.

(521-523) – Real World coordinates: Longitude and latitude coordinates of each location. Represented as an integer in meters.

Results

Distance Error (Euclidean) – mean: 9.1m / median 5.9m
Latitude Residuals – mean: 5.6m / median 3.1m
Longitude Residuals – mean: 6.1m / median 3.5m

Conclusion and Recommendations

Clearly there are problems with the fingerprints regarding specific location, to improve a future analysis it is recommended to:

- Analyze each fingerprint by unique location, remove the ones that have low signal count and select the ones that represent clearly the location.
- Analyze each cellphone and the signal received by device; there are some issues with the samples taken by some android phones.
- Analyze each user ID; there are some issues with the sampling method performed by some users.
- Combine wifi with Bluetooth technology.
- Define the WAP's by building and generate models with only the most frequent WAP's by location. This could be a good approach to predict floor and buildingID.
- Perform a PCA analysis, this will reduce the dimensionality of our dataset and improve the computational time when applying the machine learning algorithms.
- The RSSI values are in dBm (logarithmic scale), it is recommended to perform an exponential transformation to the units and check if this improves the performance of the models.
- Identify relocated WAPS.
- Perform an analysis using H20, which is a powerfull algorithm to work with large datasets.

Analysis

Pre-processing

Firstly we have to transform and clean the dataset. After revising the provided train and validation data .csv files, it was necessary to compare the names of columns of two data sets, remove repeated rows, merge two data set into one, change the values of columns, remove attributes, assign new data types, and check missing values.

Pre-process tasks:

Merge two data sets into one called “data_full”.

Repeated rows were removed (all the duplicated fingerprints).

The integer values that correspond to TIMESTAMP were changed to a class type: POSIXct.

The integer values that correspond to BUILDINGID, FLOOR, RELATIVEPOSITION, USERID and PHONEID were changed to a class type: factor.

The double values that correspond to LONGITUDE, LATITUDE, were changed to a class type: numeric.

Create an ID to which combine BUILDINGID and FLOOR

Add count of WAP's detected as feature

Data Exploration

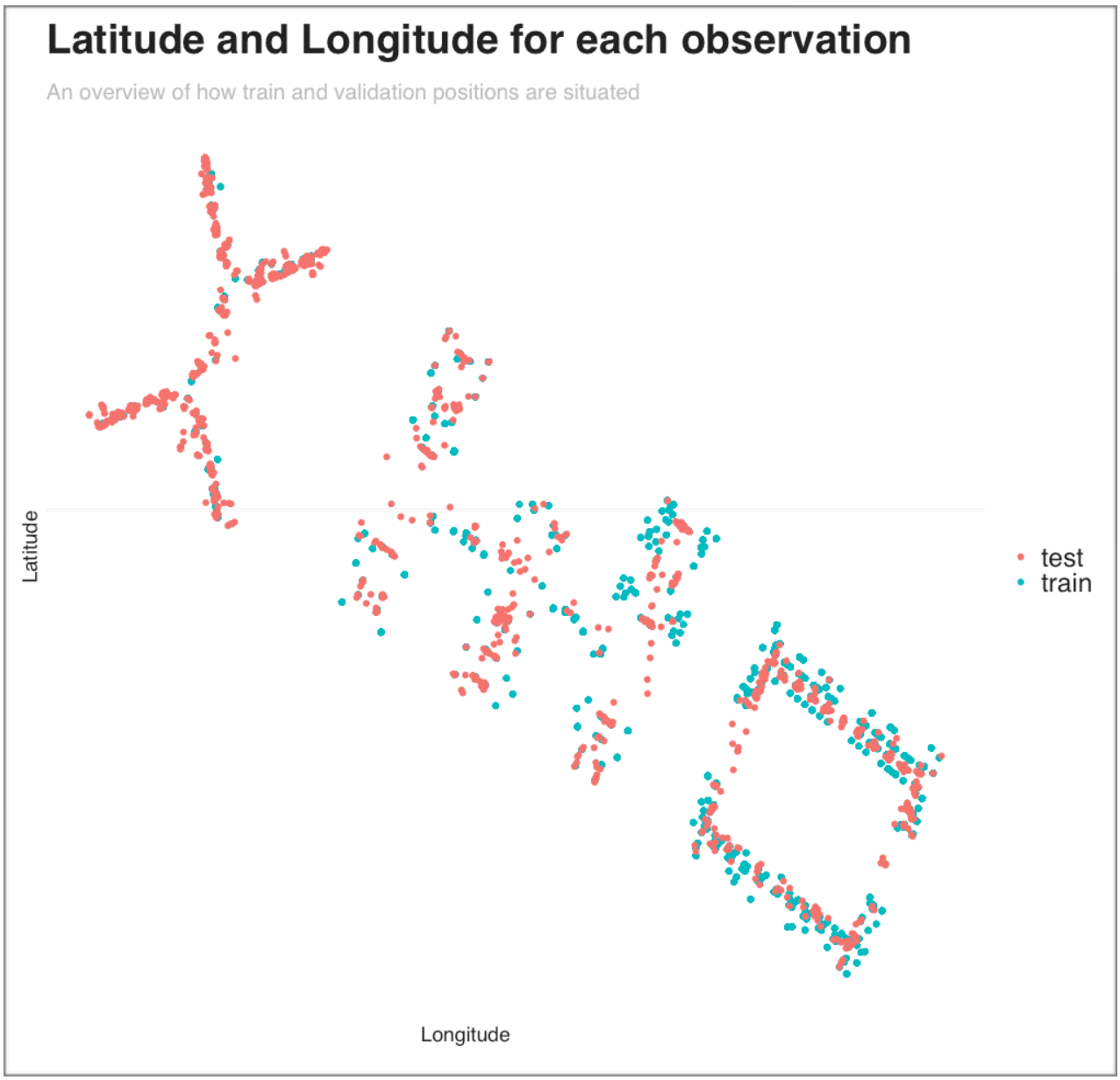

Fig 1. Distribution of each observation - Train and test sets.

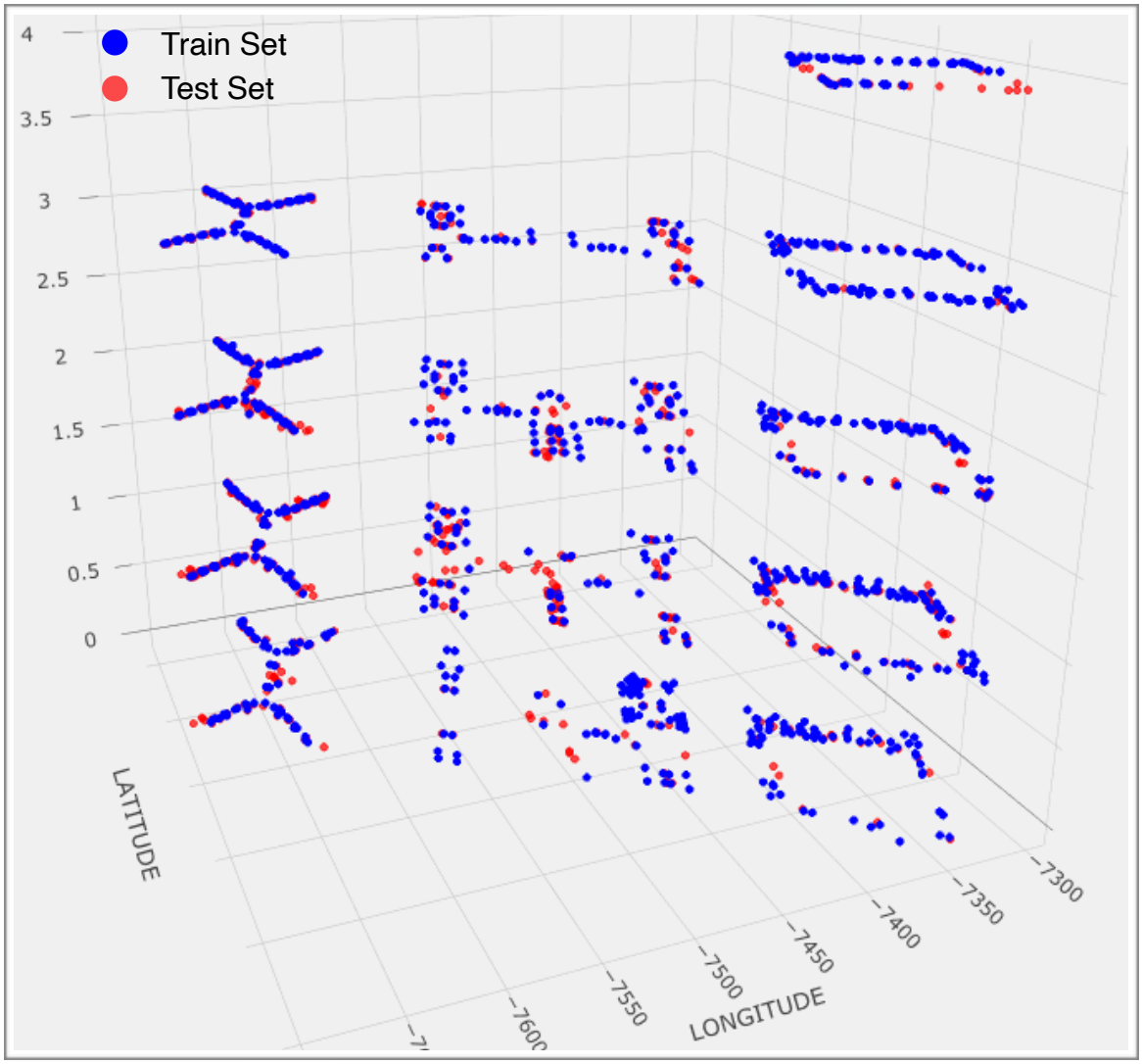

Fig 2. Distribution of each observation by building and floor - Train and test sets.

After overlaping the location of the samples in the train and the sets, some sites might present problems.

- Building 1 – Floors: 1, 2. A few measurements in the middle area and two corners.
- Building 2 – Floors: 1, 2, 4. A few measurements in the middle area.
- Building 3 – Floor: 5, Lower right corner has no measurements.

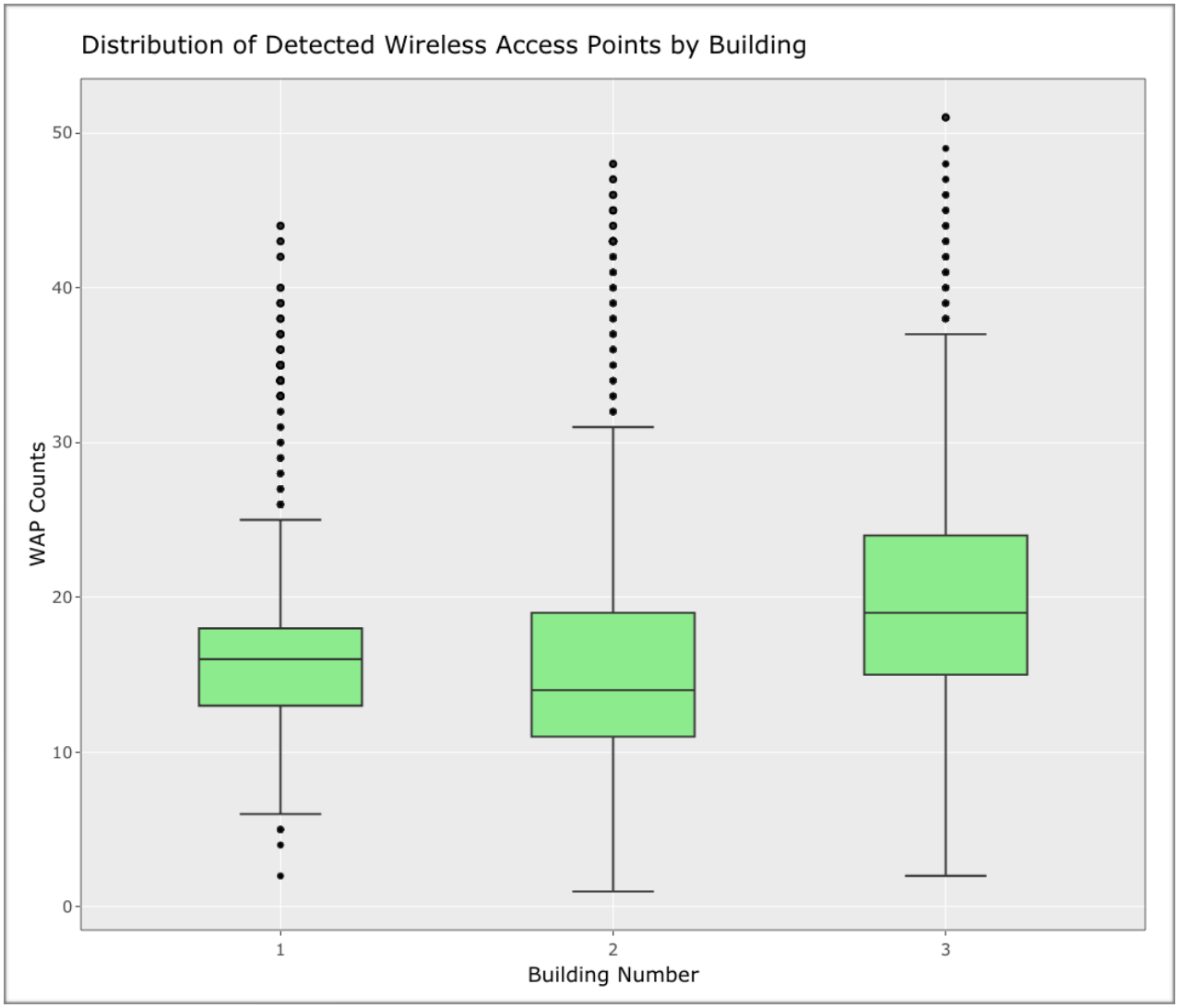

Fig 3. Histogram of distribution of detected wireless points by building

The boxplot above shows the distribution of WAPs across the buildings. Building 3 has the highest median detected WAPs whereas Building 1 and 2 appear to have similar medians. The distribution in building 1 also reaches to the lower end of WAPs detected relative to the other buildings.

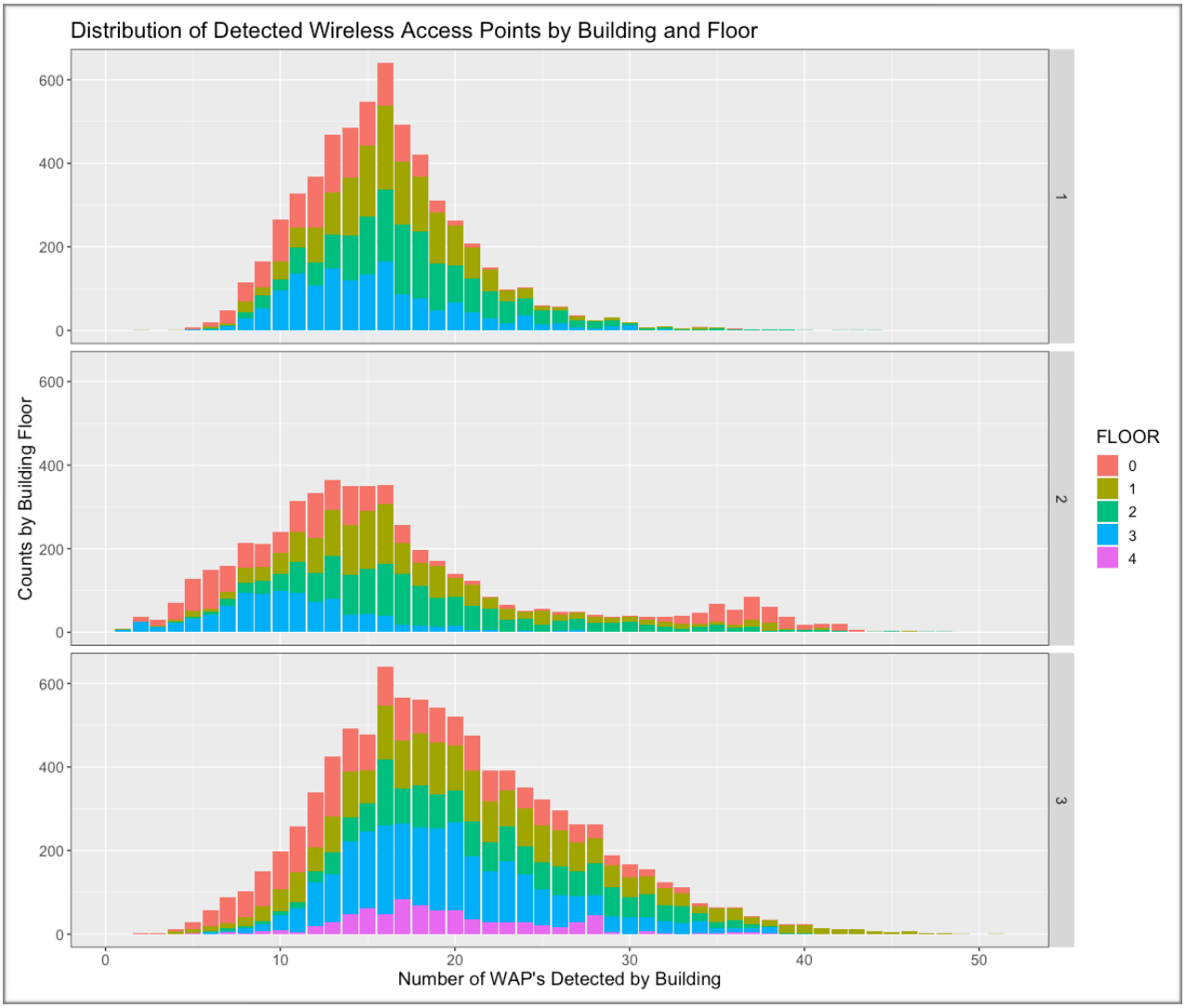

Fig 4. Histogram of distribution of detected wireless points by building and floor

The resulting histogram shows some slight differences between buildings. For one, building 3 is the only one with a 5 floors and it also has spikes in WAPs detected at 17 and 28. The distribution of WAPs detected in building 2 is more spread out than the other buildings with more occurrences of WAPs detected at the low end.

Feature Selection

It’s about creating new input features from the existing ones and selecting which columns are representative for our model to predict accurately the location

Tasks of Feature Selection:

Removing unnecessary columns

The highest WAP: we added to each row the name of the wap with the 4th highest signal.

Convert NA’s to -110 numeric value

After the selection of features, the datasets contains:

Train Data Set
520 W-Fi Access Points detected.
19.300 records used for training (fingerprints)

Test Data Set
520 W-Fi Access Points detected and 1.111 records used for validation.

Predictive Models

Building ID

Building ID Predictions

Support Vector Machine [svm]	Accuracy	0,999
	Kappa	0,999

SVM	Reference		
Prediction	1	2	3
1	536	1	0
2	0	306	0
3	0	0	268

K-Nearest Neighbor [K-NN]	Accuracy	0,997
	Kappa	0,996

K-NN	Reference		
Prediction	1	2	3
1	534	1	0
2	2	307	1
3	0	0	267

Table 1. Performance of models predicting BUILDING ID

Floor

Floor ID Predictions

Support Vector Machine [svm]

Accuracy	0,911
Kappa	0,876

SVM	Reference				
Prediction	0	1	2	3	4
0	123	24	1	0	0
1	8	421	28	0	0
2	1	15	269	7	0
3	0	2	8	164	3
4	0	0	0	1	36

Random Forrest [RF]

Accuracy	0,914
Kappa	0,880

RF	Reference				
Prediction	0	1	2	3	4
0	115	3	0	0	1
1	10	412	4	0	1
2	6	42	293	5	0
3	1	5	9	166	7
4	0	0	0	1	30

Table 2. Performance of models predicting FLOOR

Longitude

I evaluated only two (2) models (kNN and random forest) for the prediction of longitude.

	K-NN	Random Forest
RMSE	11.68	10.87
Rsquared	0.99	0.99
MAE	6.08	7.15

Table 3. Performance of models predicting LONGITUDE

We can observe that the best model to predict the longitude was a K-NN, with a tuning parameter of K= 5. In Random Forest we used a tuning parameter 100 trees and mtry= 87 (number of variables available for splitting at each tree node).

Latitude

I evaluated only two (2) models (kNN and random forest) for the prediction of longitude.

	K-NN	Random Forest
RMSE	10.36	10.42
Rsquared	0.98	0.98
MAE	5.65	6.54

Table 4. Performance of models predicting LATITUDE

We can observe that the best model to predict the latitude was a K-NN, with a tuning parameter of K= 5. In Random Forest we used a tuning parameter 100 trees and mtry= 173 (number of variables available for splitting at each tree node).

Error Analysis

Floor

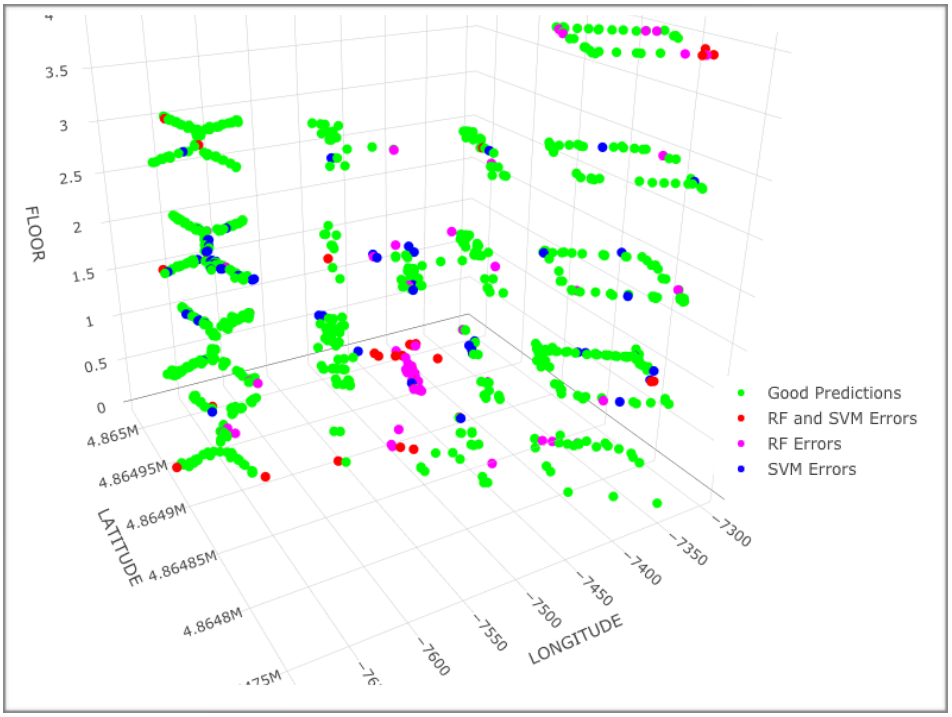

Fig 5. Location of each prediction and errors by Building and Floor

After analyzing the errors of floor predictions, we can check in detail where are the locations that are difficult to detect.

In this case, we know that there are problems with the floor 1 of building 2 and the floor 4 of building 3. We can infer some reasons for this behaviour:

- The RSSI for both locations are weak and the mobile phones are detecting WAP's in other floors or buildings.
- There are not enough "good" fingerprints for each spot, in this case the train set needs to be cleaned.
- Problems with the mobiles devices not detecting the same WAP's per location.
- The shape and location of the buildings affects the detection of WAP's, specifically in Building 2 which is located in the middle the other two buildings.

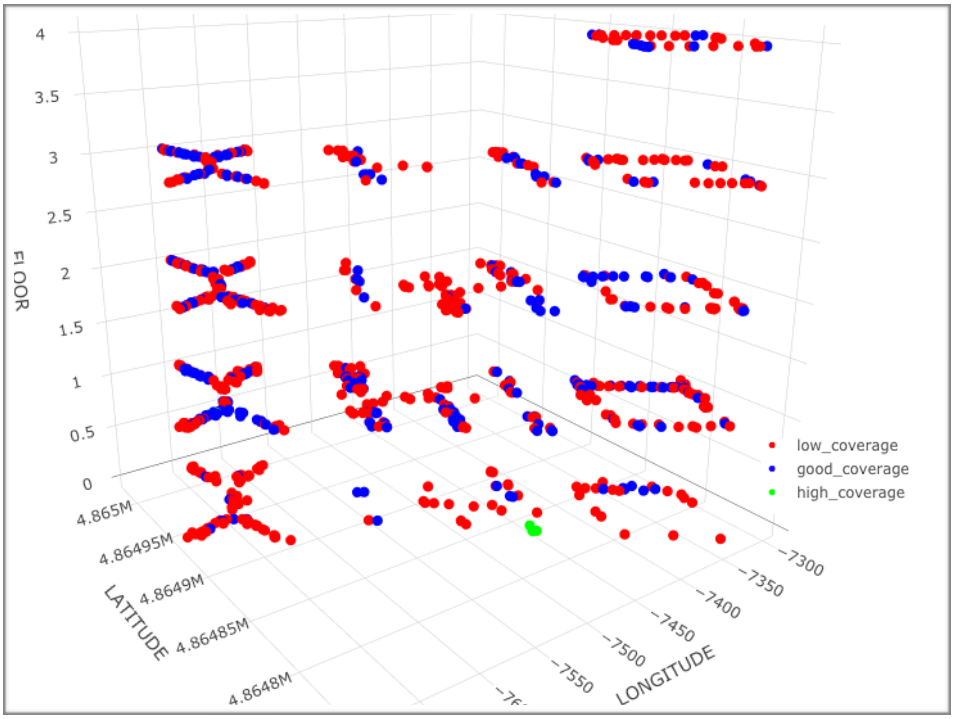

Fig 6. WAPs with different signal coverage

Longitude

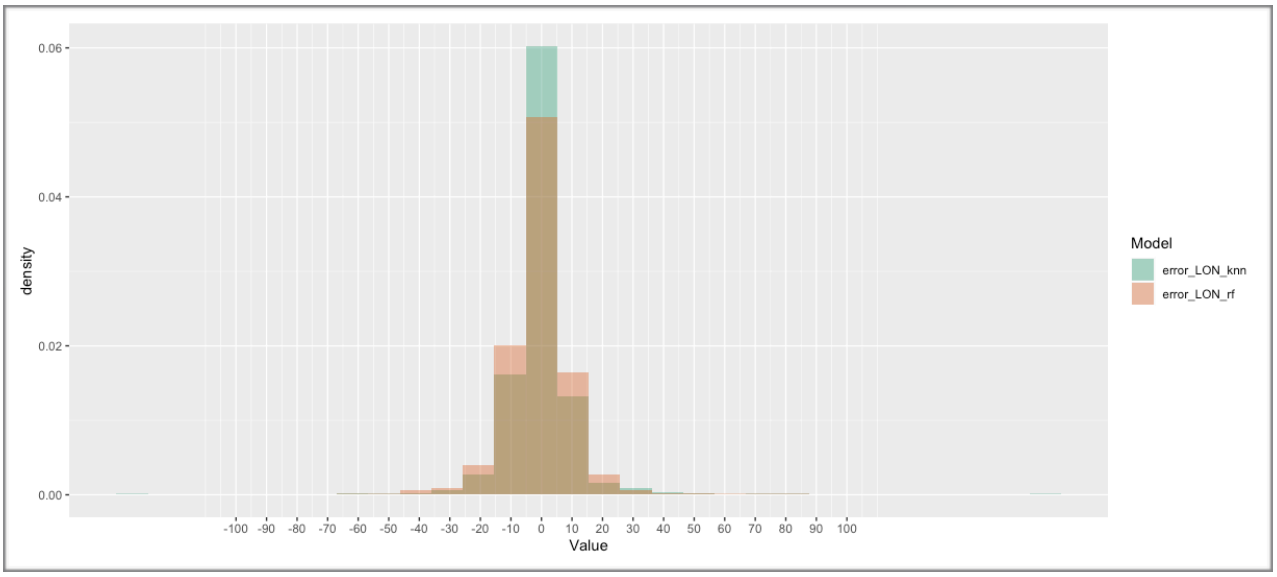

Fig 7. Comparison de error of K-NN and Random Forest in Longitude

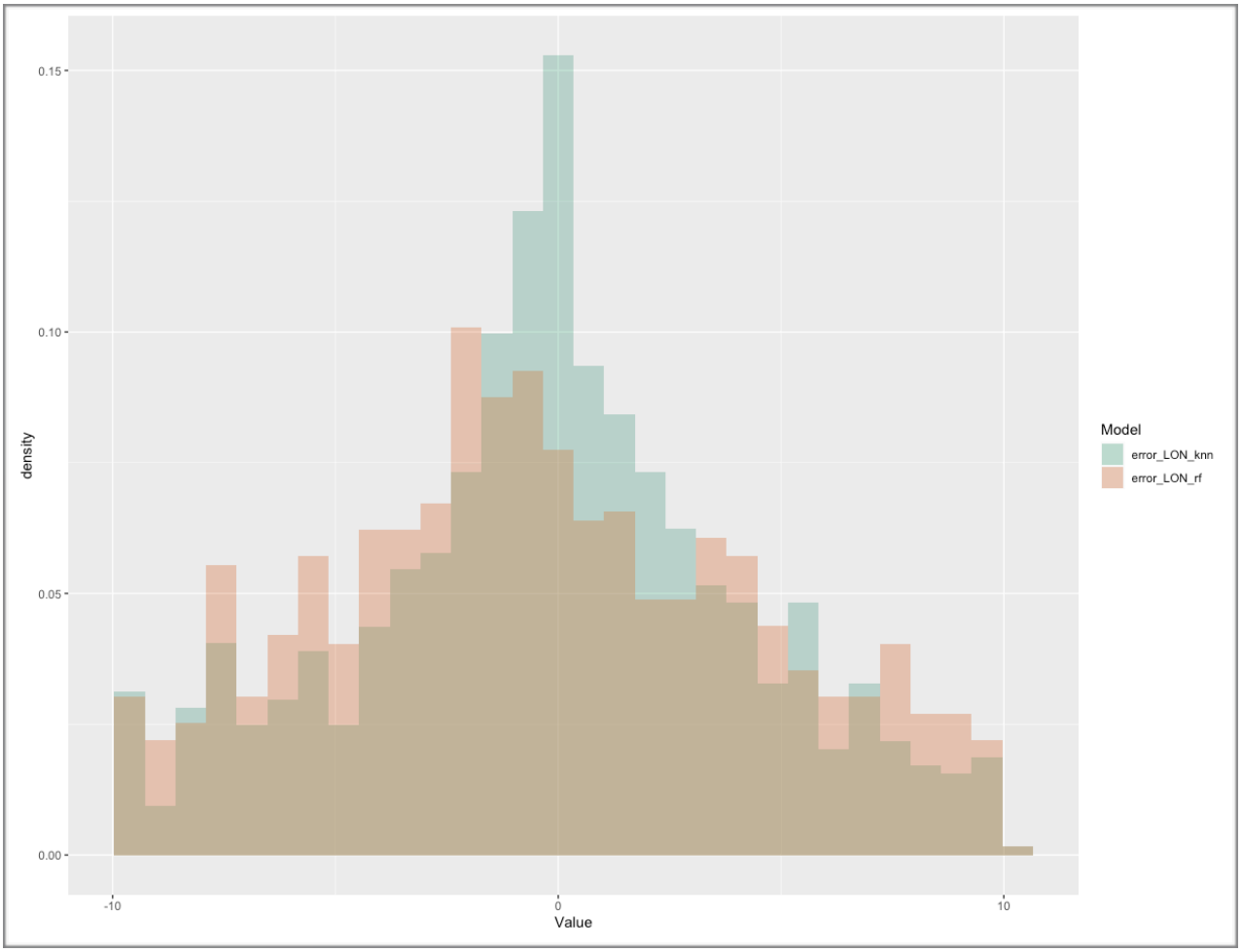

Fig 8. Comparison de error of less than 10 meters in K-NN and Random Forest in

In Fig. 7 and Fig. 8 we can see that the K-NN residuals has more density in errors of less than 10 meters, but has one outliers of 200 meters of error, meanwhile Random Forest residuals has less density in errors of less than 10 meters and some outliers of 90 meters of error.

Longitude	K-NN [mts]	Random Forest [mts]
Mean	6.08	7.15
Median	3.45	4.79

Latitude

Latitude	K-NN [mts]	Random Forest [mts]
Mean	5.65	6.55
Median	3.07	3.99

After comparing the mean and median of residuals in each model, we choose K-NN as the best model and use it to calculate the total error.

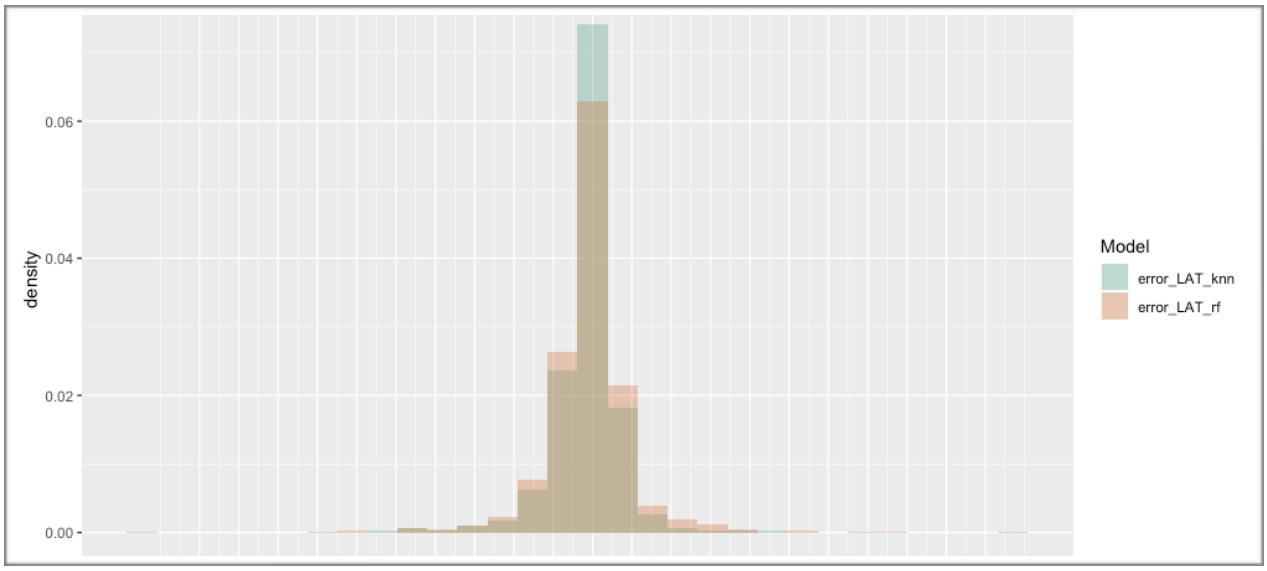

Fig 9. Comparison de error of K-NN and Random Forest in Latitude

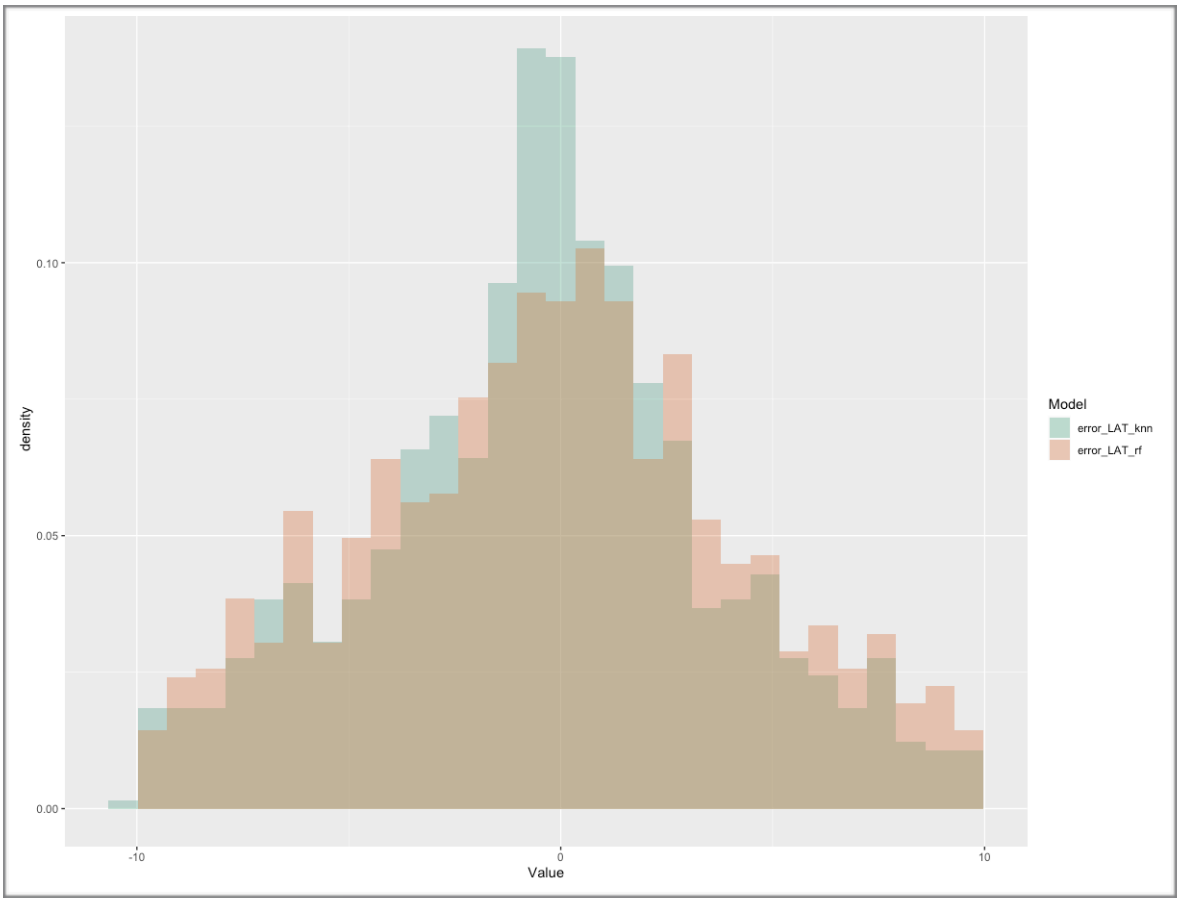

Fig 10. Comparison de error of less than 10 meters in K-NN and Random Forest in Latitude

In Fig. 9 and Fig. 10 we can see that the K-NN residuals has more density in errors of less than 10 meters, but has one outliers of 10 7 meters of error, meanwhile Random Forest residuals has less density in errors of less than 10 meters and some outliers of 80 meters of error.

Total Errors	K-NN [mts]
Mean	9.13
Median	5.89

After comparing the mean and median of residuals in each model, we choose K-NN as the best model and use it to calculate the total error.

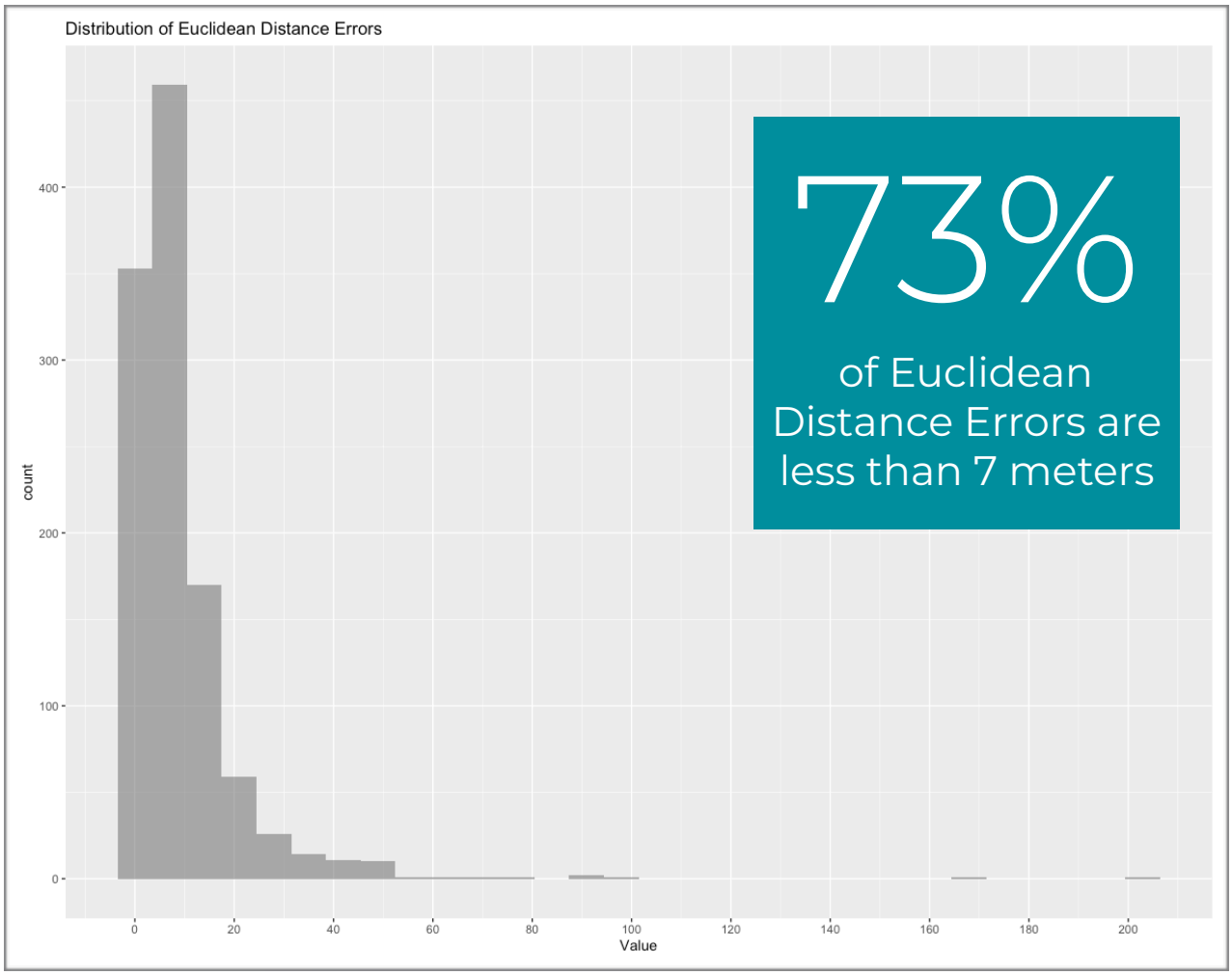

Fig 10. Distribution of Euclidean Distance of Errors