Indoor Locationing: WiFi Fingerprint

Barcelona, January 2019

Student:
Ing. Javier E. Villasmil
Villasmil@gmail.com
+58.412.2379138
+34.644.978834

INDEX

LIST OF FIGURES

LIST OF TABLES

1. EXECUTIVE 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.

In this report, we investigate the feasibility of using ***Wifi Fingerprinting*** to determine a person's location in indoor spaces.

RESULTS:

- Distance Error (Euclidean) – mean: 9m / median 7m
- Latitude Residuals – mean: 5.56m / median 3.58m
- Longitude Residuals – mean: 5.87m / median 3.98m

MODEL USED

- The best model to predict the ***BUILDINGID*** was a SVM with a tuning parameter "C" constant at a value of one. Acurracy:
    - 0.99 and Kappa: 0.99
- The best model to predict the ***FLOOR*** was a Random Forest with a tuning parameter 100 trees and 17 variables tried at each split.
    - Acurracy: 0.91 and Kappa: 0.88
- The best model to predict the ***LATITUDE*** was a Random Forest with a tuning parameter 100 trees and 104 variables tried at each split.
    - RMSE: 8.25, MAE: 5.56. R2: 0.99.
- The best model to predict the ***LONGITUDE*** was a Random Forest with a tuning parameter 100 trees and 104 variables tried at each split.
    - RMSE: 8.95, MAE: 5.87. R2: 0.99.

CONCLUSION

- 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.
    - Remove the WAP's with near zero variance; select a threshold to eliminate signals that cause noise in the model.
    - 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.
    - 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.

2. SCOPE

Investigate the feasibility of using ***Wifi Fingerprinting*** to determine a person's location in indoor spaces. Wifi fingerprinting is a method of positioning that uses the signals from multiple WiFi Access Points within a building to determine a specific location, analogously to how GPS uses satellite signals.

3. OBJECTIVES

- Analyze and understand the dataset provided by the *Universitat Jaume I* to detect possible problems regarding the values of signals for each access point and the methodology used for creating the train and validation set.
- Define the procedure used to train the models.
- Evaluate multiple machine learning models to see which produces accurate results.
- Make predictions of *Building ID*, *Floor*, *Longitude*, and *Latitude*.
- Analyze error metrics and detect in which locations the models excels or underperform.
- Elaborate recommendations on how the result might be improved.

4. 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.
    - TI-ESTCE: *Computer science and mathematics Building,* (4f).
    - TD-ESTCE: *Teacher's Building,* (4f).
    - TC-ESTCE: Science and Technology Building, (5f).
- 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.

(524) – Building ID: Integer value (from 0 to 2) that corresponds to the building in which the capture was taken.

- (0) ESTCE - TI
- (1) ESTCE - TD
- (2) ESTCE – TC

(525) – Space ID: Contains a single integer value that is used to identify the particular space (offices, labs, etc.) where the capture was taken.

(526) – Relative Position: Relative position with respect to the space ID. It denotes if the capture was taken inside (integer value 1) or outside (integer value 2) the space.

(527) – User Identifier: Integer value that ranges from one (1) to eighteen (18). This value represents the eighteen (18) different users who participated in the procedure to generate the training samples.

(528) – Phone Identifier: Integer value that ranges from zero (0) to twenty-four (24). This represents the Android devices used in each capture. Each Phone ID it is associated to an Android cellphone model and version.

(529) – Timestamp: Integer representing the time (UNIX time format) in which each capture was taken.

5. ANALYSIS

5.1 PRE-PROCESSING

The first step to start analyzing the data was to transform and clean the dataset. After revising the provided *.csv* file, it was necessary to change the values of columns, remove attributes, assign new data types, and check missing values.

The actions taken to pre-process the data were:

- The integer values that correspond to the ***Building ID*** were changed to a ***class type: factor*** with three (3) levels. Each level was renamed according to the buildings real identity.
- The integer values that correspond to the ***Floor*** were changed to a ***class type: factor*** with three (5) levels – one for each floor.
- ***UNIX time*** was converted into ***class type: POSIXct / POSIXt*** for dates and times.
- Repeated rows were removed (all the duplicated fingerprints).
- Wireless Access Points (WAP's) without RSSI measurements in any of the registers were removed in both sets (train and validation)
- Registers without RSSI measurements were removed in both sets (train and validation)
- RSSI values with no signal, i.e., ***+100dBm*** were change to ***-105dBm***, a value below the lowest signal detected. This will improve the model's accuracy.
- The data set was complete and ***there were not missing values***.

5.2 DATA EXPLORATION

To investigate the location of each sample a 3D plot was elaborated (Fig. 1), by doing this it is possible detect if a specific area has less measurements that another. This could cause a bias in the models by lowering the accuracy when trying to predict certain spaces.

Equally, the same approach was used to analyze the validation set (Fig. 2).

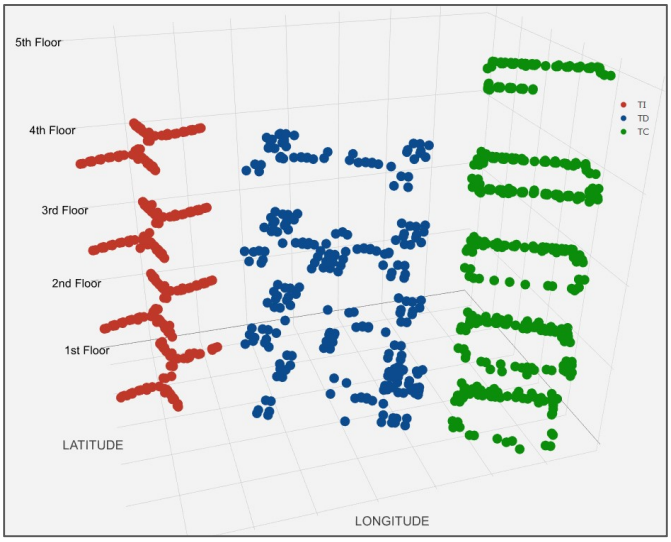

Fig 1. Distribution of samples for each Building ID and Floor – TRAIN SET.

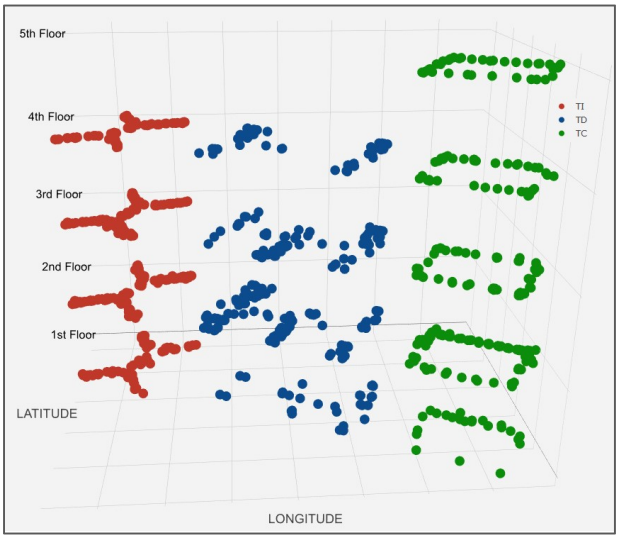

Fig 2. Distribution of samples for each Building ID and Floor – TEST SET.

After comparing the location of the samples in the train set against the validation set, four (4) sites might present problems.

- **Building TC-03** – 5th Floor: Lower right corner has no measurements.
- **Building TD-02** – 1nd Floor: Few measurements in the middle area.
- **Building TD-02** – 2th Floor: Few measurements in the middle area.
- **Building TD-02** – 4th Floor: Few measurements in the middle area.

Moreover, it is possible to overlap both dataset (train and validation) to observe clearly the spaces with lower measurements.

In this case, how the locations are dispersed by floor. See Fig. 3, Fig 4 and Fig. 5.

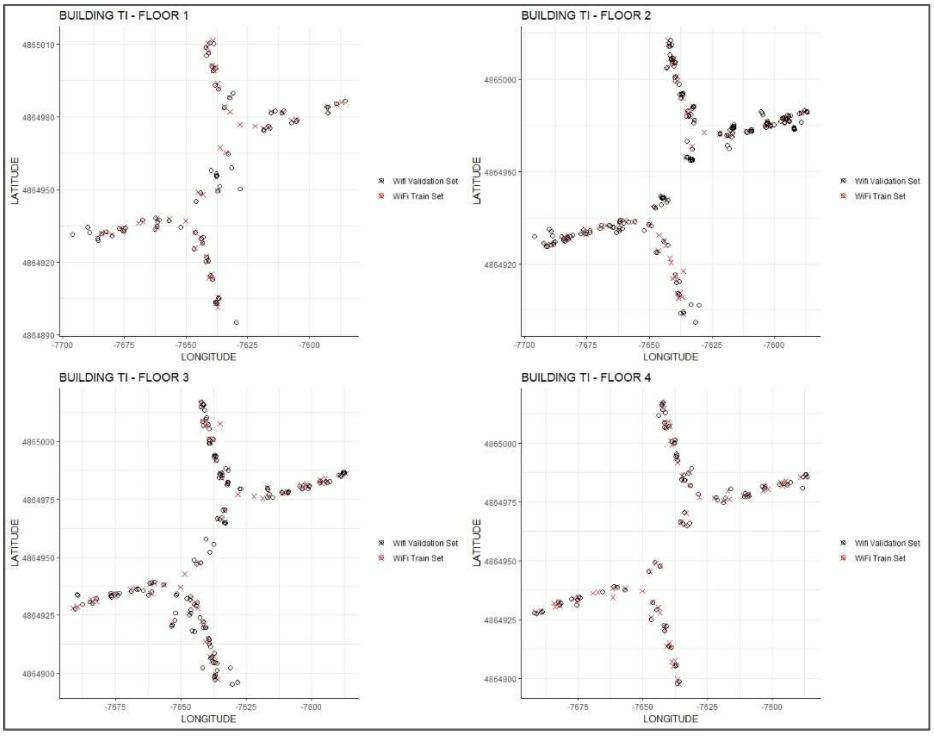

Fig 3. Train and Validation points in the building TI-00 by floor.

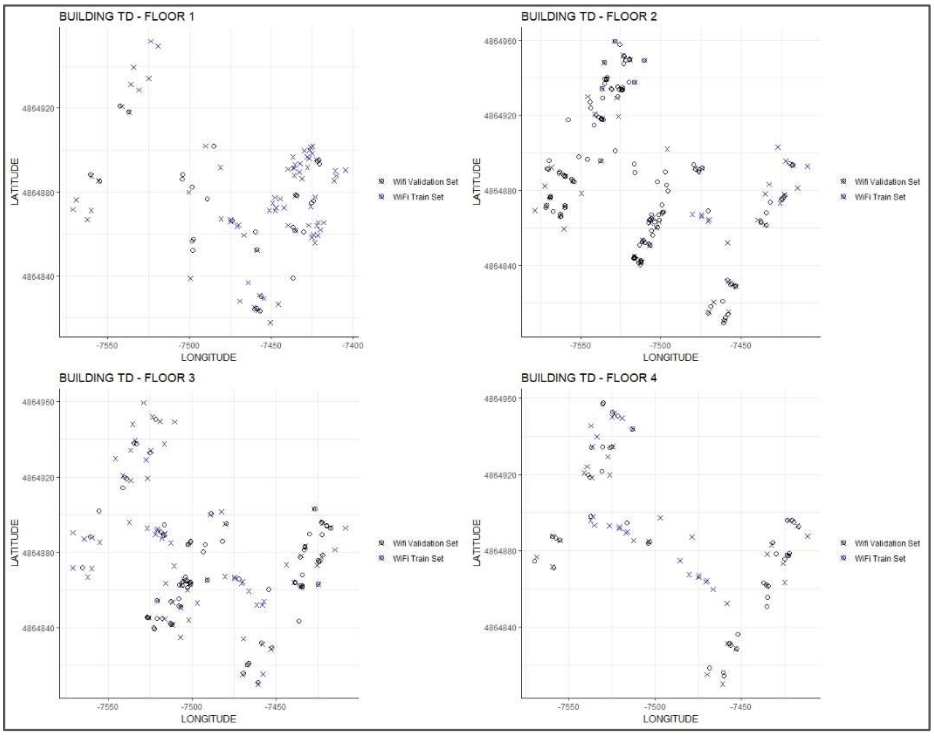

Fig 4. Train and Validation points in the building TD-01 by floor.

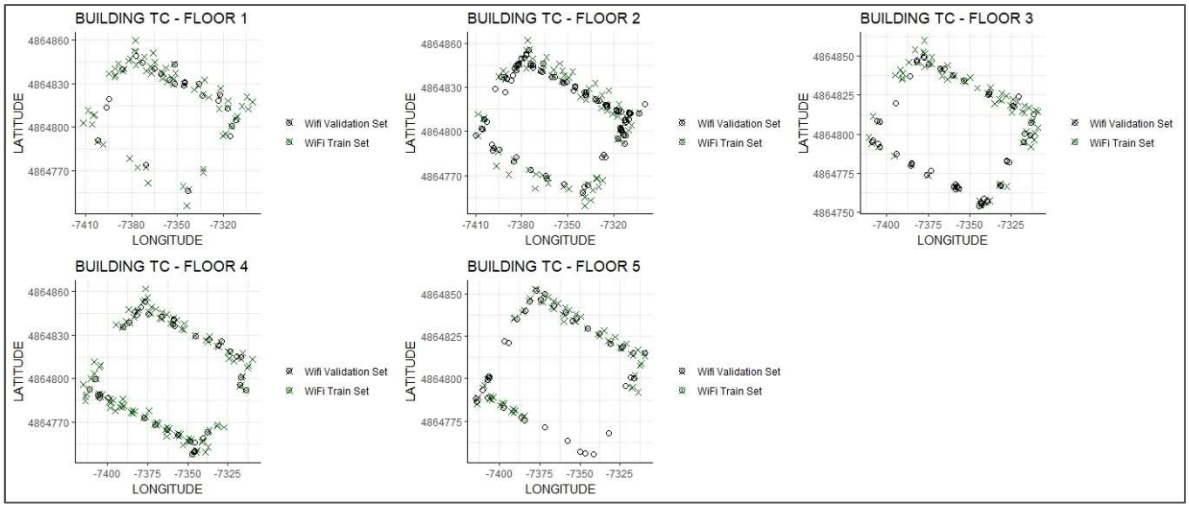

Fig 5. Train and Validation points in the building TC-02 by floor.

In addition, by checking the frequencies of the sampling it is clear that the building ***TC-02*** has more fingerprints but badly distributed on the 5th floor, while the ***TI-00*** and ***TD-01*** buildings have similar number of samples better distributed. See fig. 6

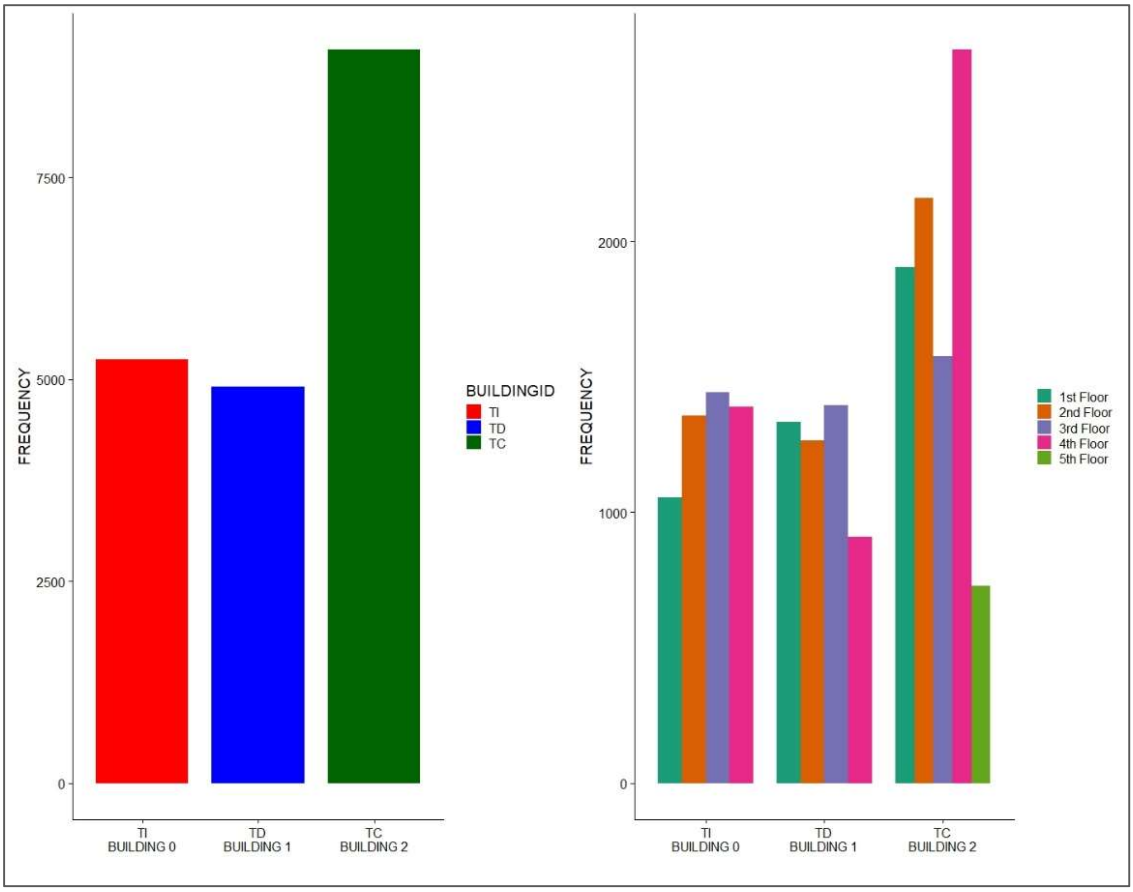

Fig 6. Frequencies of registers (fingerprints) by building and floor.

For this analysis, we neglected the variables relative position, user identifier, phone identifier and time.

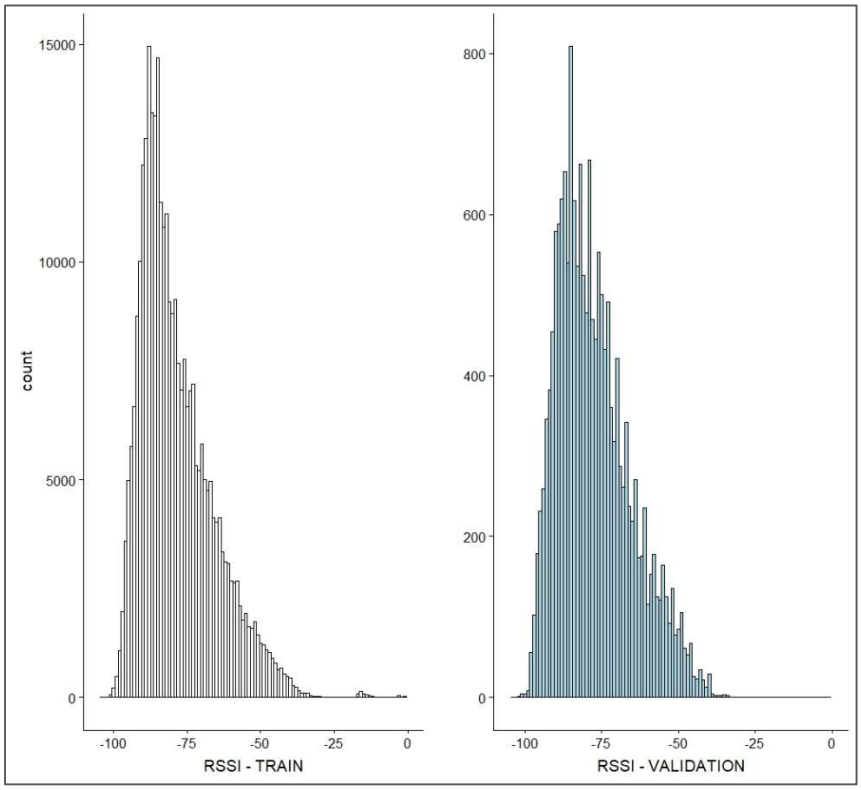

Fig 7. Frequencies of RSSI – TRAIN and VALIDATION.

Furthermore, it is important to compare the frequency of signals values from the train and validation set. We can see in Fig. 6 that both distributions are similar, and ranges from -100dBm to -30dBm.

In the train set there are values between -25dBm and 0dBm, these signals are not accurate because they are physically hard to achieve, probably a faulty receiver (mobile phone) is generating these values.

5.3 FEATURE SELECTION

Selecting which columns are representative for our model to predict accurately the location of a person is a task where the analyst needs to compare different relationships between your dependent variable and predictors.

In this case, since the method proposed is “fingerprinting” we decided to only remove the zero-variance columns in both datasets (train and test), in this way, we maintain the integrity of the data without removing WAP’s with few signals that might improve our model.

Following the same procedure, for each row duplicates were removed to prevent redundancy in fingerprints, leaving duplicates could generate a bias in the models towards a specific result.

After the selection of features, the datasets contains:

- TRAIN
  - 312 W-Fi Access Points detected, 208 were removed.
  - 19.226 records used for training (fingerprints). 711 were removed.
- TEST
  - 312 W-Fi Access Points detected, 208 were removed.
  - 1.111 records used for validation.

5.4 METHODOLOGY

To predict the location of a person in a building four (4) variables needs to be predicted: ***BuildingID***, ***Floor***, ***Latitude*** and ***Longitude***.

After evaluating each of these variables, a mixed approach was selected. First, a model to predict the ***building*** is trained and applied, and then with the predictions of this model we train three new models for the remaining three variables.

In this case, we are using a *"cascade model"* with **buildingID** and a *"parallel model"* with **floor**, **latitude** and **longitude**.

For each interation, three algorithms were used and compared: SVM, kNN and Random Forest.

5.5 PREDICTIVE MODELS

5.5.1 BUILDINGID

	kNN	Random Forest	✔ SVM
Accuracy	0.9936	0.9972	0.9990
Kappa	0.9900	0.9985	0.9985

Table 1. Performance of different models predicting BUILDING ID.

We can observe that the best model to predict the building was a ***SVM*** with a tuning parameter *"C"* constant at a value of one. See table 1.

In addition, the confusion matrix of the SVM model presents only one miss classification (See table. 2); therefore, we can conclude that these building predictions can be used as an input for the other models (cascade approach).

		REFERENCE		
		TI-00	TD-01	TC-02
PREDICTION	TI-00	535	0	0
	TD-01	1	307	0
	TC-02	0	0	268

Table 2. Confusion Matrix for BuildingID – SMV model.

5.5.2 FLOOR

	kNN	✔ Random Forest	SVM
Accuracy	0.8118	0.9108	0.8775
Kappa	0.7394	0.8752	0.8300

Table 3. Performance of models predicting FLOOR.

We can observe that the best model to predict the floor was a ***Random Forest*** with a tuning parameter ***100 trees*** and ***17 variables*** tried at each split. See table 3.

In addition, the confusion matrix of the Random Forest presents several points of misclassification (See table. 4).

For example, the models predicts a 3rd floor when the real value is in a 2nd or 1st floor, also some locations are predicted two or three floors away of the reference point, this behavior needs to be checked.

PREDICTION \ REFERENCE	1st Floor	2nd Floor	3rd Floor	4th Floor	5th Floor
1st Floor	116	3	0	0	1
2nd Floor	9	412	6	0	1
3rd Floor	6	42	291	6	0
4th Floor	1	5	9	165	9
5th Floor	0	0	0	1	29

Table 4. Confusion Matrix for FLOOR – Random forest model.

Since this confusion matrix groups the results of the three (3) buildings, we do not know if the accuracy of the model over one building is better than another, to check this this we applied the model to a subset of the validation set, selecting each building separetly.

This allow us to compare performances and metrics in each building using the same model that was trained with the entire dataset. See table 5, table 6 and, table 7.

PREDICTION \ REFERENCE	1st Floor	2nd Floor	3rd Floor	4th Floor	5th Floor
1st Floor	72	2	0	0	0
2nd Floor	3	204	3	0	0
3rd Floor	3	2	161	2	0
4th Floor	0	0	1	83	0
5th Floor	0	0	0	0	0

Accuracy:	0.9701
Kappa:	0.9578

Table 5. Confusion Matrix and performance by each FLOOR – BUILDING TI-00.

PREDICTION \ REFERENCE	1st Floor	2nd Floor	3rd Floor	4th Floor	5th Floor
1st Floor	23	1	0	0	0
2nd Floor	3	98	1	0	0
3rd Floor	3	39	82	4	0
4th Floor	1	5	4	43	0
5th Floor	0	0	0	0	0

Accuracy:	0.8013
Kappa:	0.7133

Table 6. Confusion Matrix and performance by each FLOOR – BUILDING TD-01.

PREDICTION \ REFERENCE	1st Floor	2nd Floor	3rd Floor	4th Floor	5th Floor
1st Floor	21	0	0	0	1
2nd Floor	3	109	2	0	0
3rd Floor	0	2	47	0	0
4th Floor	0	0	5	39	9
5th Floor	0	0	0	1	29

Accuracy:	0.9142
Kappa:	0.8830

Table 7. Confusion Matrix and performance by each FLOOR – BUILDING TC-02.

After analyzing the confusion matrixes for each building, we can observe that our model is accurate predicting the floors in two buildings: *TI-00* with 97% accuracy and *TC-02* with 92% accuracy.

In addition, we noticed that building *TD-01* is the one that has problems with locations between the 2nd and 3rd floor (80% accuracy), this causes a decrease in the general accuracy of our random forest model.

5.5.3 LATITUDE

For the prediction of *latitude* we evaluated only two (2) models (kNN and random forest) since we noticed a constant under performance of the SVC algorithm compared to the other two.

	kNN	✔ Random Forest	SVM
RMSE	15.83	8.25	x
MAE	7.56	5.57	x
R^2	0.95	0.99	x

Table 8. Performance of models predicting LATITUDE.

We can observe that the best model to predict the Latitude was a ***Random Forest*** with a tuning parameter ***100 trees*** and ***104 variables*** tried at each split. See table 8.

These metrics groups the errors of the three (3) buildings; to check each building separately we applied the model to a subset of the validation set. See Table 9.

Random Forest - LATITUDE

	BUILDING TI-00	BUILDING TD-01	BUILDING TC-02
RMSE	5.77	10.60	9.24
MAE	3.99	7.35	6.42
R^2	0.97	0.91	0.90

Table 9. Error metrics by building using the random forest model for predicting LATITUDE.

After analyzing the performance of the model for each building, we notice that the LATITUDE is predicted with a MAE of 4 meters in the building TI-00, while in the buildings TD-01 and TC-02 the MAE is between 6 and 7 meters.

5.5.4 LONGITUDE

For the prediction of *longitude* we evaluated only two (2) models (kNN and random forest) since we noticed a constant under performance of the SVC algorithm compared to the other two.

	kNN	✔ Random Forest	SVM
RMSE	19.71	8.95	x
MAE	7.89	5.88	x
R^2	0.97	0.99	x

Table 10. Performance of models predicting LONGITUDE.

We can observe that the best model to predict the longitude was a ***Random Forest*** with a tuning parameter ***100 trees*** and ***104 variables*** tried at each split. See table 11.

These metrics groups the errors of the three (3) buildings; to check each building separately we applied the model to a subset of the validation set. See Table 9.

Random Forest - LONGITUDE

	BUILDING TI-00	BUILDING TD-01	BUILDING TC-02
RMSE	6.84	8.98	10.93
MAE	4.73	6.44	7.30
R^2	0.94	0.96	0.88

Table 11. Error metrics by building using the random forest model for predicting LONGITUDE.

After analyzing the performance of the model for each building, we notice that the LONGITUDE is predicted with a MAE of 4.73 meters in the building TI-00, while in the buildings TD-01 and TC-02 the MAE is between 9 and 11 meters.

5.6 ERROR ANALYSIS

5.6.1 FLOOR

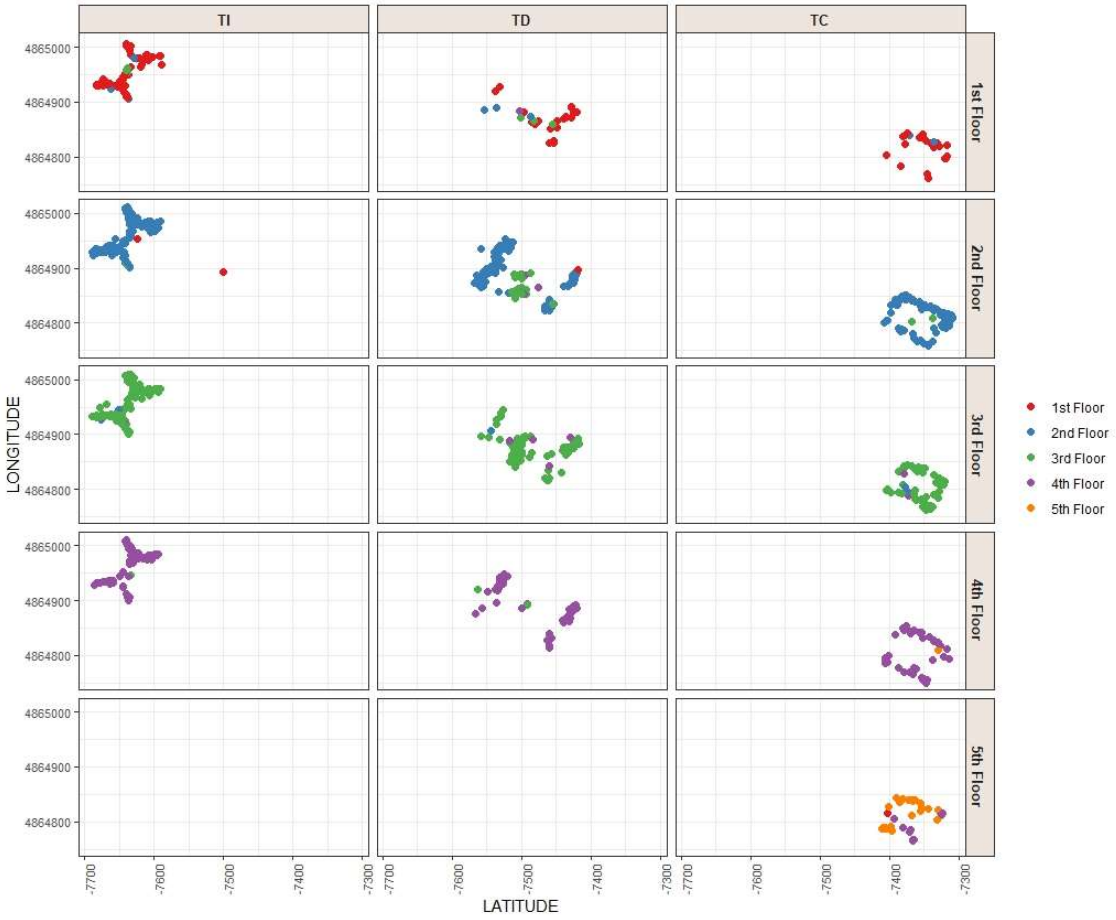

Fig 8. Location of each prediction by BUILDING and FLOOR.

After analyzing the confusion matrix of the 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 *2nd floor of building TD-01* and the *5th floor of building TC-02.*

We can infer three (4) reasons for this behavior:

- 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 *TD-01* which is located in the middle the other two buildings.

5.6.2 LATITUDE

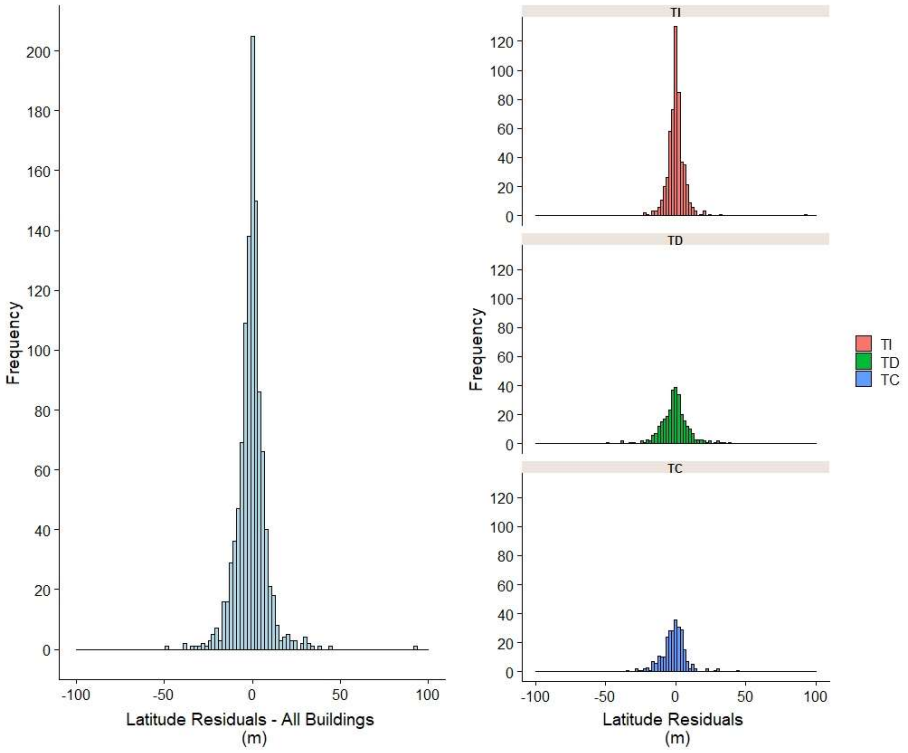

Fig 9. Latitude – residuals histogram.

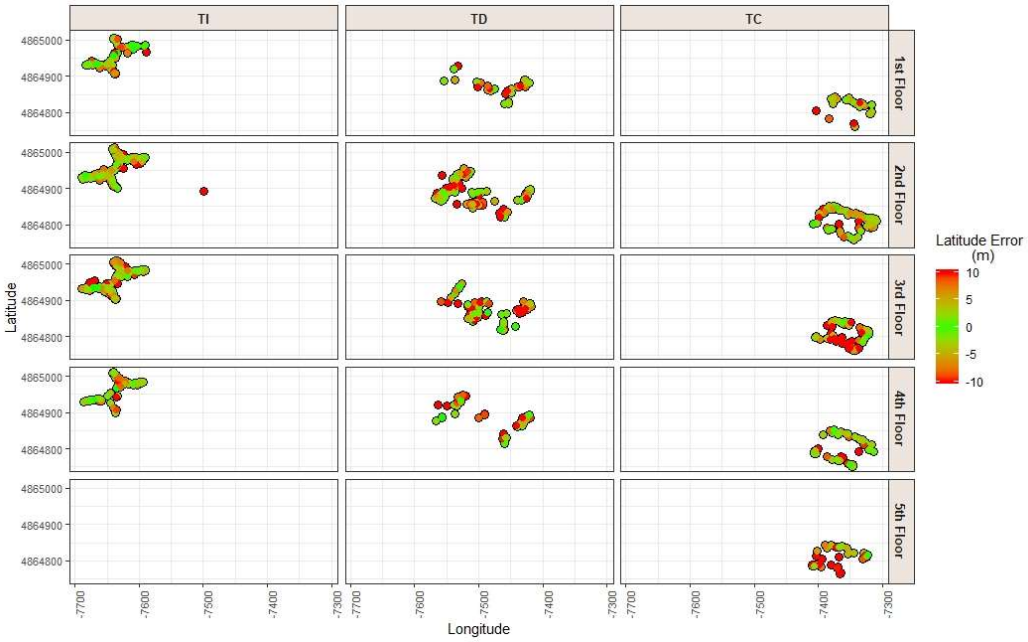

Fig 10. Latitude – residuals grid.

We can observe in fig. 9 that the residuals of latitude are distributed close to zero; there is one outlier in the building *TD-00* cause by misclassification when predicting the *building ID*.

In general, the mean error between the reference and the prediction when predicting the LATITUDE is **5.56*m*** and the median ***3.58m***.

In addition, the 3rd and 5th floor of the building *TC-02* have larger residuals, meaning that the predictions are far from the reference point.

5.6.3 LONGITUDE

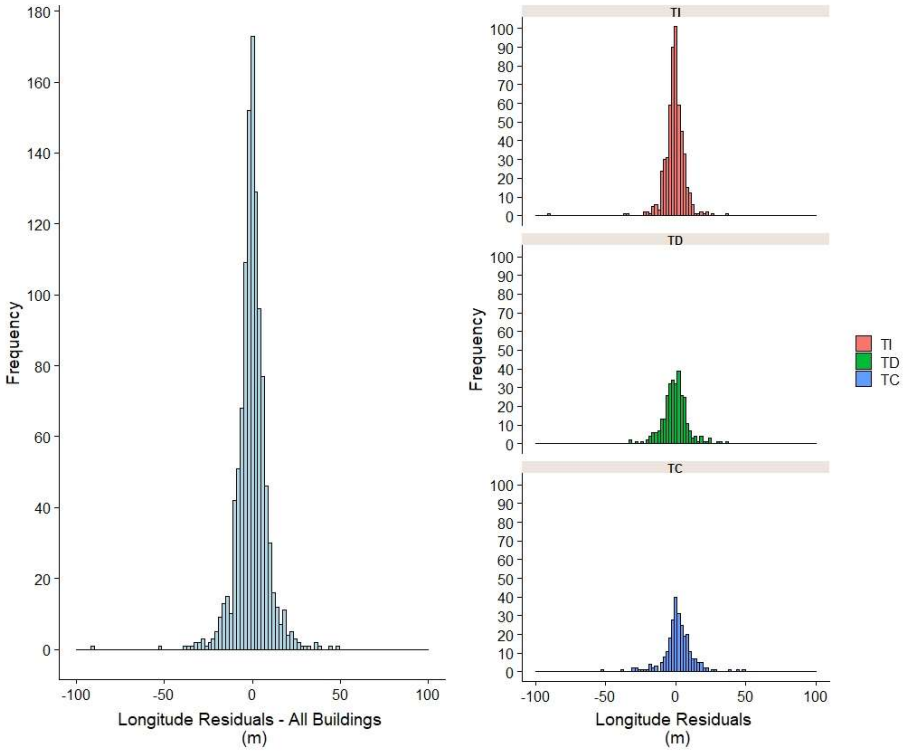

Fig 11. Longitude – residuals histogram.

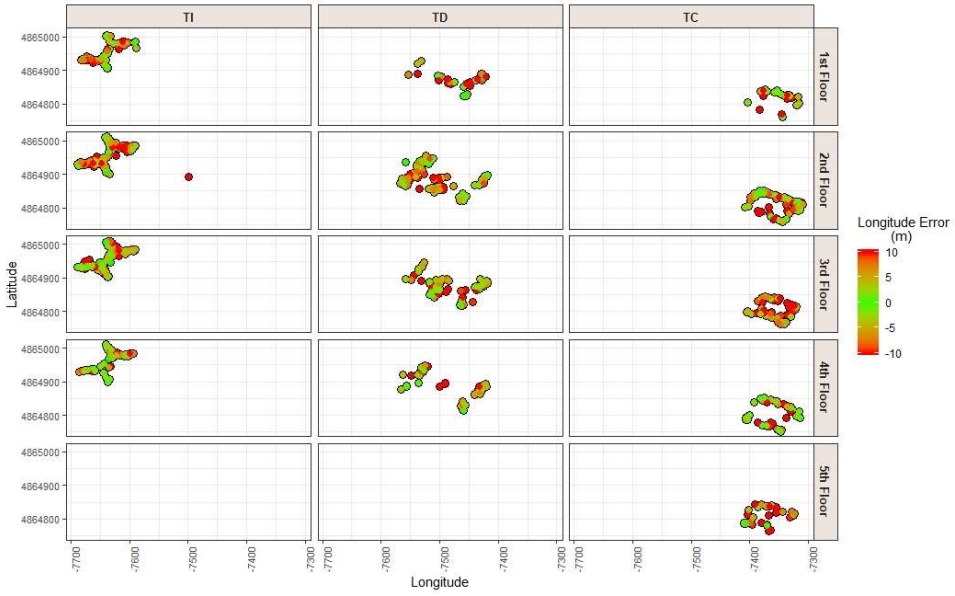

Fig 12. Longitude – residuals grid.

We can observe in fig. 11 that the residuals of longitude are distributed close to zero; there is one outlier in the building *TD-00* cause by misclassification when predicting the *building ID*.

In addition, the 3rd and 5th floor of the building *TC-02* and the 2nd and 3rd floor of the building *TD-01* have larger residuals, meaning that the predictions are far from the reference point.

In general, the mean error between the reference and the prediction when predicting the LONGITUDE is **5.87*m*** and the median ***3.98m***.

5.6.4 DISTANCE

To calculate the distance between the reference point and the predicted location the Euclidean distance was computed using the following formula:

$$\boldsymbol{Distance\_Error} = \sqrt{(\boldsymbol{latitude_{real}} - \boldsymbol{latitude_{predicted}})^2 + (\boldsymbol{longitude_{real}} - \boldsymbol{longitude_{predicted}})^2}$$

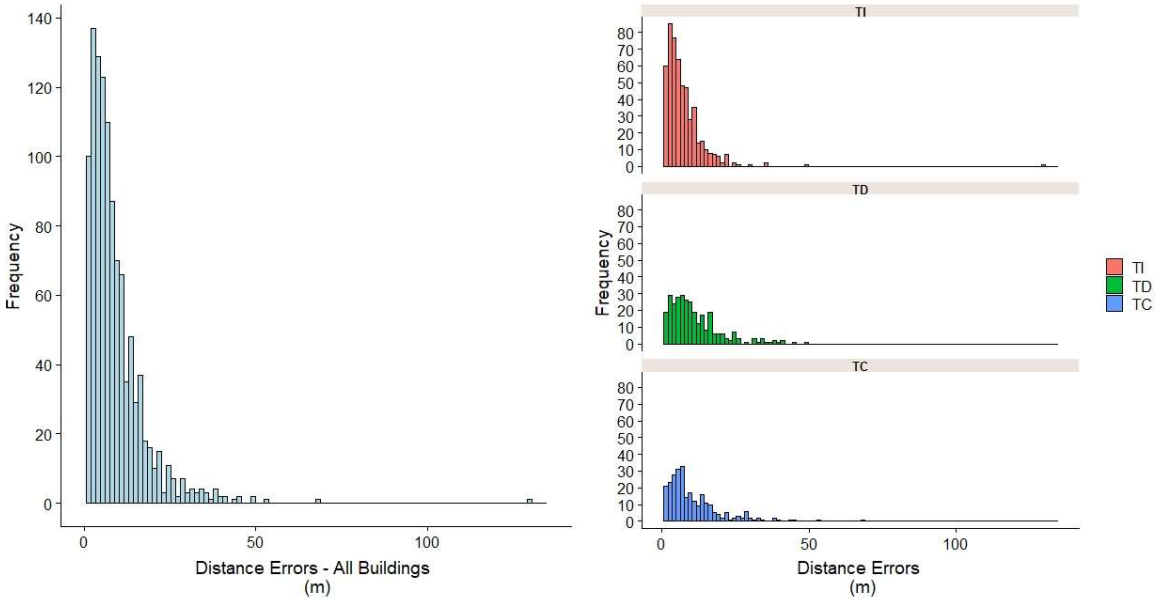

Fig 13. Distance error – histogram.

We can observe in fig. 13 that the distance errors ranges from 0 to 60 meters; there is one outlier in the building *TI-00* cause by misclassification (over 100 meters of error) when predicting the *building ID*.

In general, the mean error between the reference and the prediction is ***9m*** and the median ***7m***.

In addition, the building *TI-00* is the one with the best predictions, while the buildings TD-01 and TC-02 have values of error distances above 30 meters that needs to be checked.

In the fig. 14 we can see the boxplot plot for each building.

The distribution of distance errors are closer to zero in the building *TI-00* and the outliers are smaller, while in the *TD-01* and *TC-02* there are more point scattered around higher distances errors.

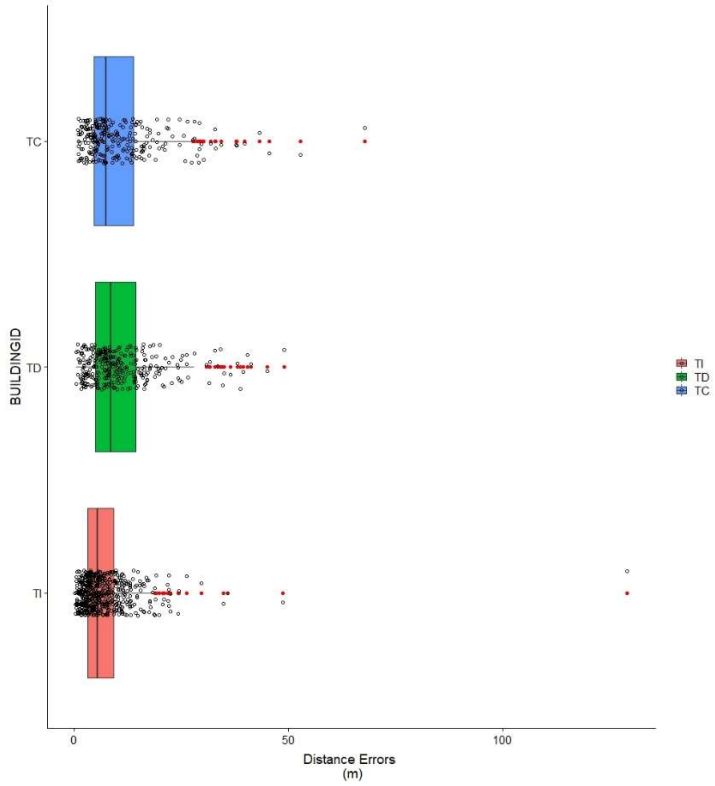

Fig 14. Distance error –Boxplot.

To understand the cause of the errors and the location of each, a vector field is plotted. By doing this we can determine the direction of a prediction with respect to the reference point.

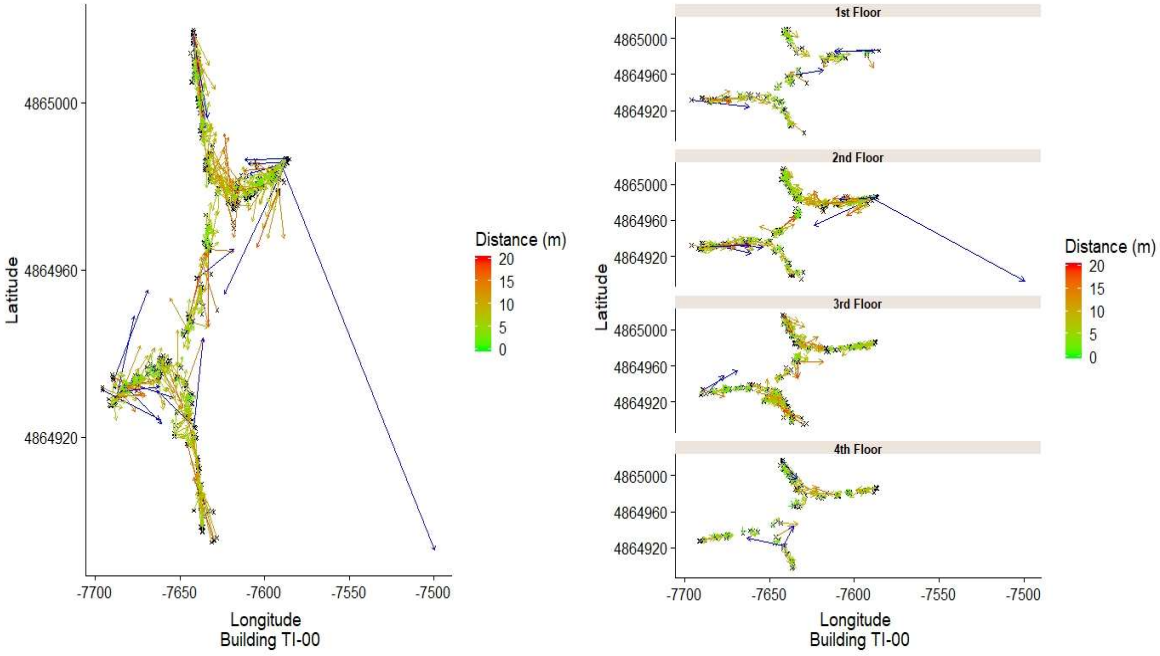

Fig 15. Vector Field – Building TI-00.

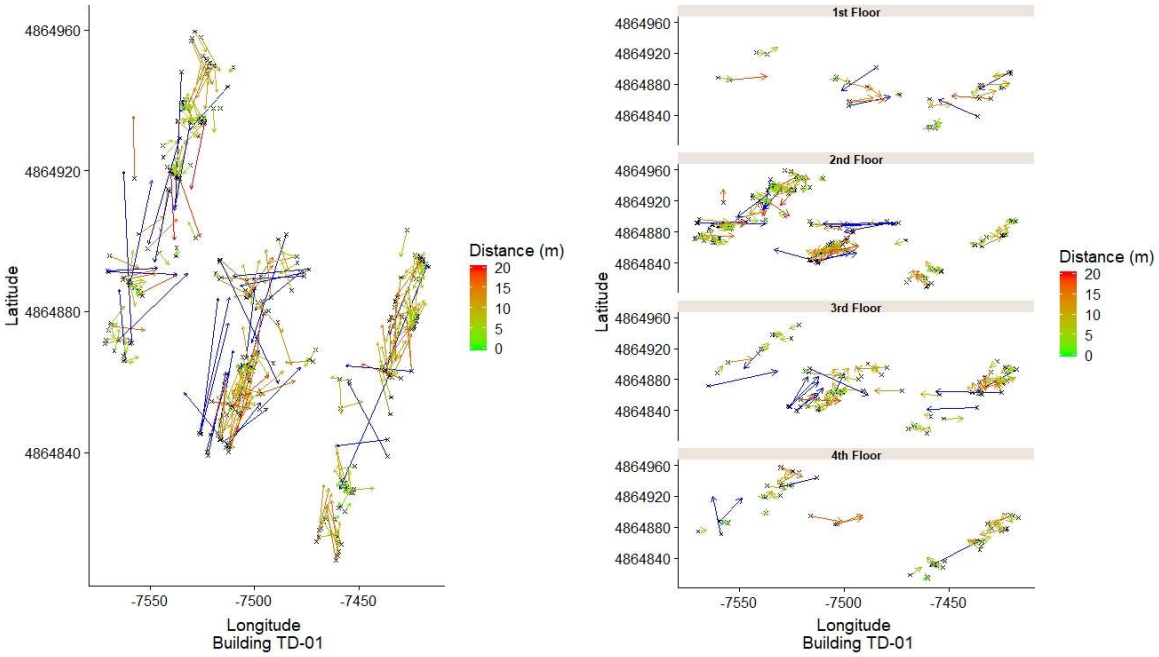

Fig 16. Vector Field – Building TD-01

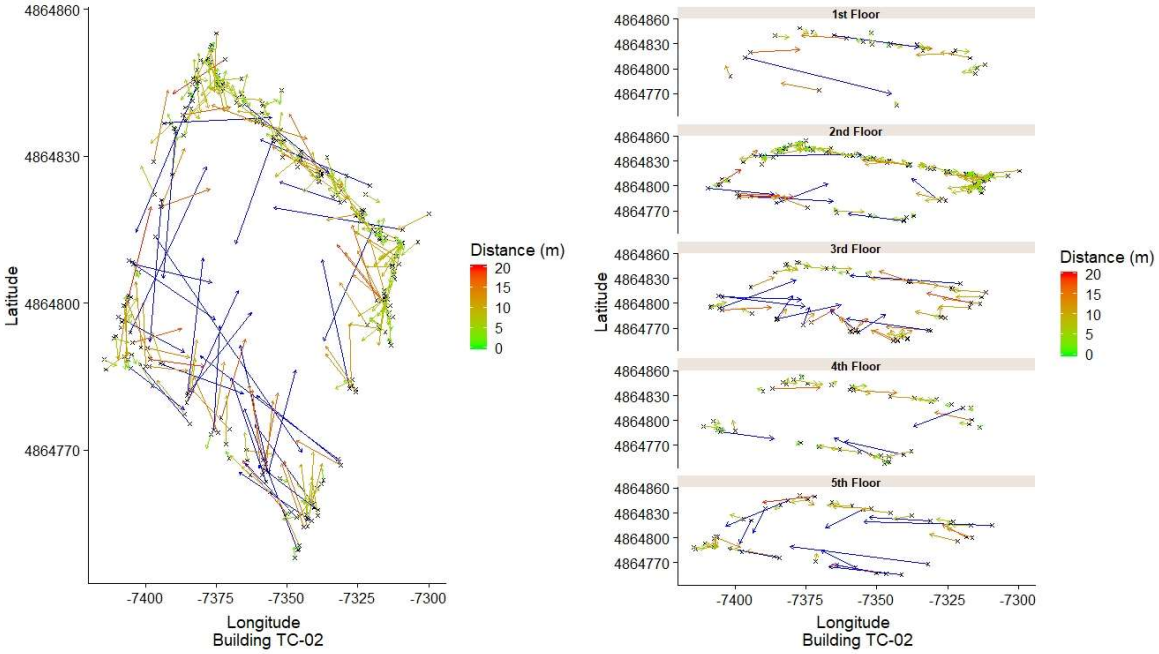

Fig 17. Vector Field – Building TC-02

In the fig 15, we observe for the building *TI-00* that the longest distances between the references point and the predictions are located in the southwest and northeast corner; this could be because the fingerprints in these points are detecting WAP's located in other buildings.

In addition, the performance of the model in this building *(TI-00)* is good with almost all predictions getting located near the reference point by 10 meters or less.

Furthermore, in the fig 16, we observe for the building *TD-01* that the longest distances between the references point and the predictions are located in the central and west area.

In general, the shape of the building and the small hallways tends to blocks the signal reception of the devices; this generates bad fingerprints that affects the model in the training phase.

Also, since *TD-01* is between the other two buildings the interference and WAP's detections affects the sampling of good fingerprints.

Finally, in the fig 17, we observe that for the building *TC-02* the longest distances between the references point and the predictions are located in the south wing.

Notice that the fingerprints located in the north wing of the buildings are clearly stronger and with better quality, this cause that the other locations tends to be predicted towards this side; this generates a bias in the model since the predictions are getting located in a space outside the building.

6. CONCLUSIONS

- The dataset was pre-processed using the following procedure:
  - Removed the zero-variance columns in both datasets (train and test).
  - Removed rows duplicates to prevent redundancy in fingerprints.
  - Removed registers (rows) without RSSI measurements.
  - The integer values that correspond to the ***Building ID and Floor*** were changed to a ***class type: factor***
  - ***UNIX time*** was converted into ***class type: POSIXct / POSIXt*** for dates and times.
  - RSSI values with no signal, i.e., ***+100dBm*** were change to ***-105dBm***.

- The data set was complete and ***there were not missing values***.

After the selection of features, the datasets contains:

  - TRAIN
    - 312 W-Fi Access Points detected, 208 were removed.
    - 19.226 records used for training (fingerprints). 711 were removed.
  - TEST
    - 312 W-Fi Access Points detected, 208 were removed.
    - 1.111 records used for validation.

- The methodology selected to perform the analysis was a "cascade model" with ***buildingID*** and a "parallel model" with ***floor***, ***latitude*** and ***longitude***. For each iteration, three algorithms were used and compared: ***SVM***, ***kNN*** and **Random Forest.**
  - The best model to predict the ***BUILDINGID*** was a ***SVM*** with a tuning parameter *"C"* constant at a value of one. *Acurracy: 0.99* and *Kappa: 0.99*
  - The best model to predict the ***FLOOR*** was a ***Random Forest*** with a tuning parameter ***100 trees*** and ***17 variables*** tried at each split. *Acurracy: 0.91* and *Kappa: 0.88*
  - The best model to predict the ***LATITUDE*** was a ***Random Forest*** with a tuning parameter ***100 trees*** and ***104 variables*** tried at each split. *RMSE: 8.25, MAE: 5.56.* R^2*: 0.99.*
  - The best model to predict the ***LONGITUDE*** was a ***Random Forest*** with a tuning parameter ***100 trees*** and ***104 variables*** tried at each split. *RMSE: 8.95, MAE: 5.87.* R^2*: 0.99.*

- The mean error between the reference and the prediction when predicting the ***LATITUDE*** is **5.56m** and the median **3.58m.**

- The mean error between the reference and the prediction when predicting the ***LONGITUDE*** is **5.87*m*** and the median error ***3.98m***.
- The mean error between the reference and the prediction (Euclidean distance between points) is ***9m*** and the median ***7m***.

- Clearly there are problems with the fingerprints regarding two buildings: *TD-01 and TD-02*, 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.
    - Remove the WAP's with near zero variance; select a threshold to eliminate signals that cause noise in the model.
    - 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.
    - 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.