Indoor Positioning System Report

Wasabees

2023-12-08

1. Executive Summary

This project summarizes the findings and recommendations of an indoor positioning system (IPS) using Wi-Fi strength signals.

The primary objective is to implement a robust model that can accurately predict the location of a device connected to the local Wi-Fi network. We implemented two supervised machine learning methods (XGBoost and K-Nearest Neighbors) to create two separate models. Notably, both models show predictive skills with a mean error under 2.5 meters for K-Nearest Neighbors, and 1.7 for XGBoost. Not only XGBoost demonstrated to have better accuracy in the predictions, but has also the advantage of being computationally efficient. Further accuracy can be obtained by optimizing the location of the routers inside the building to ensure an even distribution of the signals. This recommendation is easy and cost-effective to implement.

2. Introduction and Background

Indoor position systems (IPS) development is an active area of research with multiple applications ranging from medical settings to security. Problems associated with tracking indoor location include low accuracy from traditional GPS systems due to problems in signal propagation[1]. An alternative comes from using Wi-Fi signal strength to triangulate the position of a device. The existence of numerous public and private Wi-Fi networks in buildings, makes this approach highly practical.

Limitations to this approach include the decreasing strength as the device moves away from the source, the high sensitivity of the orientation of the device to the reception of Wi-Fi signals[2], and noise introduced by other access points. Other facotrs include building layouts and construction material.

Specifically for orientation, some positions allow for a direct line of connectivity between the router and the receiver, while others, the user's body might form an obstruction[3].

The problem to solve to develop an IPS can be summarized as the following question: "Given a set of signatures (signal strength) from nearby Wi-Fi routers at a given device orinetation, and considering the decay of the distance with signal, what's the location?.

This type of problem, in which the desired outcome, in this case location, is explicitly stated as some part of the dataset, allows the use of supervised machine learning (SML). SML to put it simply, takes features from the data (variables believe to have a relationship with the response), to predict the answer. In this case, we use features such as signal strength and orientation to predict positions x, y in the grid. To validate and test a model the data is typically divided into "training data" and "testing data". Training data is used to develop a model that can predict location and then, its accuracy with unseen data is assessed with the testing data. A high accuracy in the training data and low on the testing means that the model is not able to find general patterns applicable to other observations. This level of accuracy can also be understood as one that minimizes the error between the values predicted by the model and the real ones.The average difference of this error is known as Mean Average Error (MAE), which we use to validate our model. Because this metric is measuring the difference between the predicted values and the real values, a good model will have a low MAE.

3. Data

The following section, describes and characterizes the data set provided by the client.

The client gridded 540 m^2 of their building into 166, 1-meter-by-1-meter cells in which measurements of signal strength from 6 access points (Wi-Fi routers) were obtained from a handheld device connected to a local Wi-Fi network (Fig. 1). The data are subdivided into two sub-sets, “offline” and “online”, distinguished by the fact that the offline data set was sampled at fixed locations and orientations, versus the online data set that was sampled at random locations and orientations.

The offline data, intended to train a model, was collected at all the 166 fixed points within the grid. At each location, the device was oriented in 45-degree increments, starting from 0 up to 360 degrees, resulting in 8 angles (i.e., 0, 45, 90, 135, 180, 225, 270, and 315). Signal strength for the access points was measured for each orientation a total of 110 times, totaling 880 samples per location and a total of 146,080 observations.

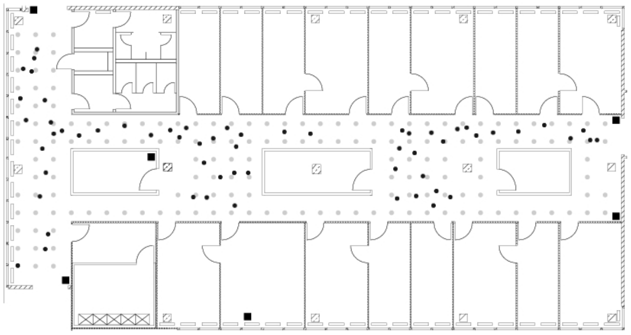

Figure 1: Floor plan of the client’s building where the data was collected. Dots mark the locations for the data sampling. Black dots correspond to the offline locations, and grey dots correspond to the online data locations. Black squares are the 6 access points in the building. Note the gridded area is confined to locations outside the rooms.

The online data was designed to simulate real-world data, where devices are held at random orientations and are not bound to the center of a grid point. 60 locations were randomly selected and the device was then oriented at a random angle. Similarly, to the offline data set, signal strength for the access points was measured 110 times, resulting in a total of 6600 observations in total.

The data is composed by 8 variables from which we remove those with redundant or non-relevant information after the data exploration and only keep:

1) Orientation: The scanning device’s orientation at binned 45 degrees increments.
2) Mac: The IP address of the access points.
3) Signal: Signal strength in decibel-milliwatts.
4) PosX and PosY: x and y coordenates in the grid.

The data exploration also reveals an inverse relationship between signal strength and distance to an access point (Fig 2), and the influence of orienttion on signal strength (Fig. 3).

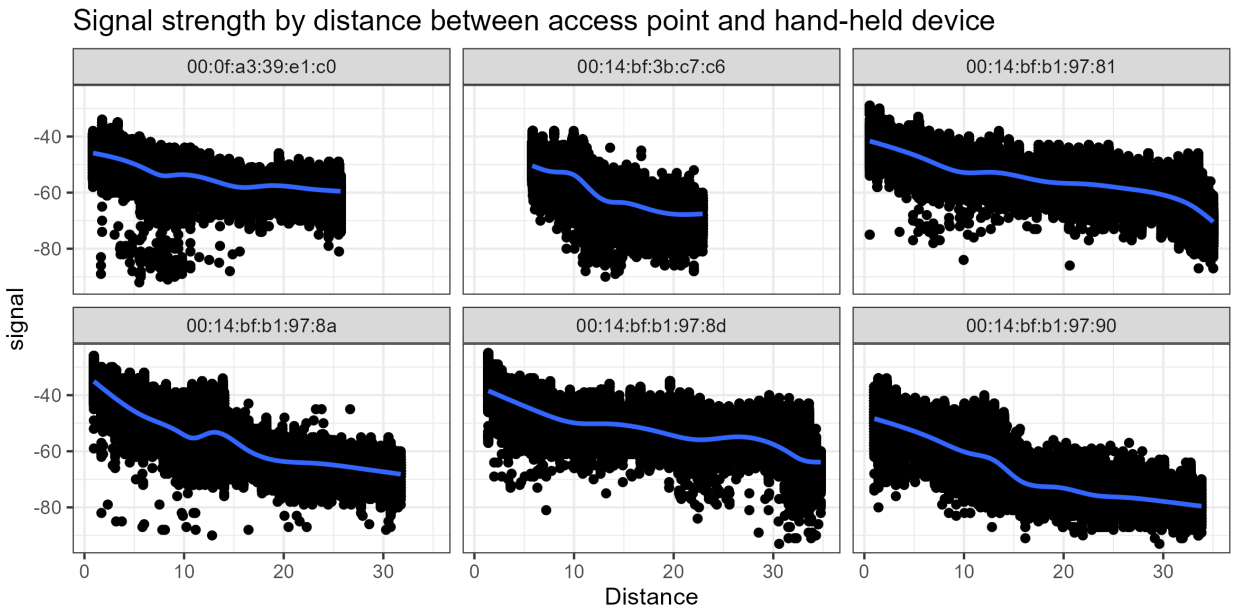

Figure 2: Scatter-plot of signal vs distance grouped by access point. Signal strength is inversely related to distance. Some access point have consistent lower signal strengths

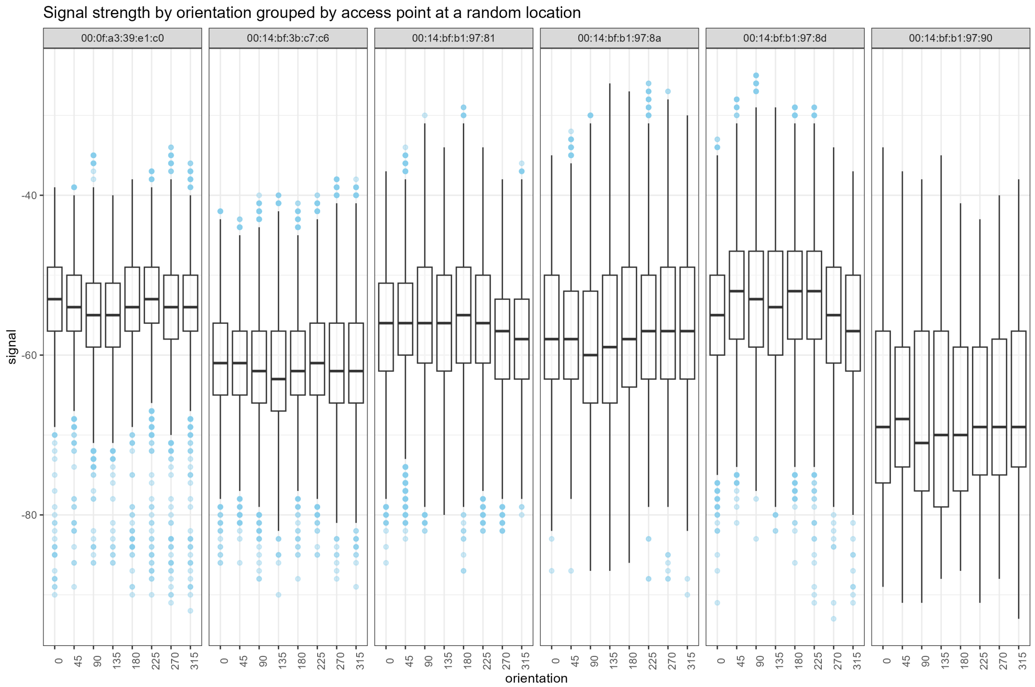

Figure 3: Box-plot of signal strength depending on orientation, grouped by access point at a random location

4. Methods

We develop the IPS using two supervised machine learning (ML) algorithms: eXtreme Gradient Boosting (XGBoost) and K-Nearest Neighbors (KNN). In supervised machine learning, algorithms are trained with labeled examples to find patterns in the input data. To put it simply, we train models that take values for strength signal, orientation and relative position to access points to predict a location in the grid. Each location in the data set is characterized by a combination of values for strength signal and device orientation that can be used to train a model able to find and generalize these patterns. In order to train a model that can find and then generalize these patterns, we use the mean values for signal strength on each orientation on the offline data set.

An important aspect of supervised machine learning is reserving part of the data to validate and quantify the accuracy of the models, in this case, we utilize the totality of the offline data to train the model, and the online to test it.

The following sections briefly describe the algorithms utilized.

4.1 K-Nearest Neighbors

This approach, in broad terms, asigns a value for latitude and longitude based on the similarity of the testing point to others in the training data set. Specifically, K in K-Nearest Neighbors refers to the number “k” of similar data points. This similarity or proximity is evaluated based on the unique characteristics of the neighbors. For example, we would expect a data point located at the margins of the building to have on average a weaker signal strength creating a unique fingerprint for each location. Although this algorithm can be highly accurate, is also sensitive to noisy data and its computational costs increase rapidly in large data sets.

4.2 XGBoost

XGBoost is a very powerful ML algorithm that creates predicted values by creating a series of decision trees[4]. This means that the final predicted value is a combination of the values found per tree. One of the advantages of using XGBoost in this study is the capacity of handling large data sets and high accuracy.

5. Results

Both models created by XGBoost (Fig. 4) and KNN (Fig. 5) show predictive skills. Both approaches show more robust predictions on the left side of the floor plan. The errors in the predictions for XGBoost and KNN are represented by black lines in the figures below are on average 1.7 meters and 2.5 meters (Appendix A) demonstrating the high accuracy of both methods to predict location.

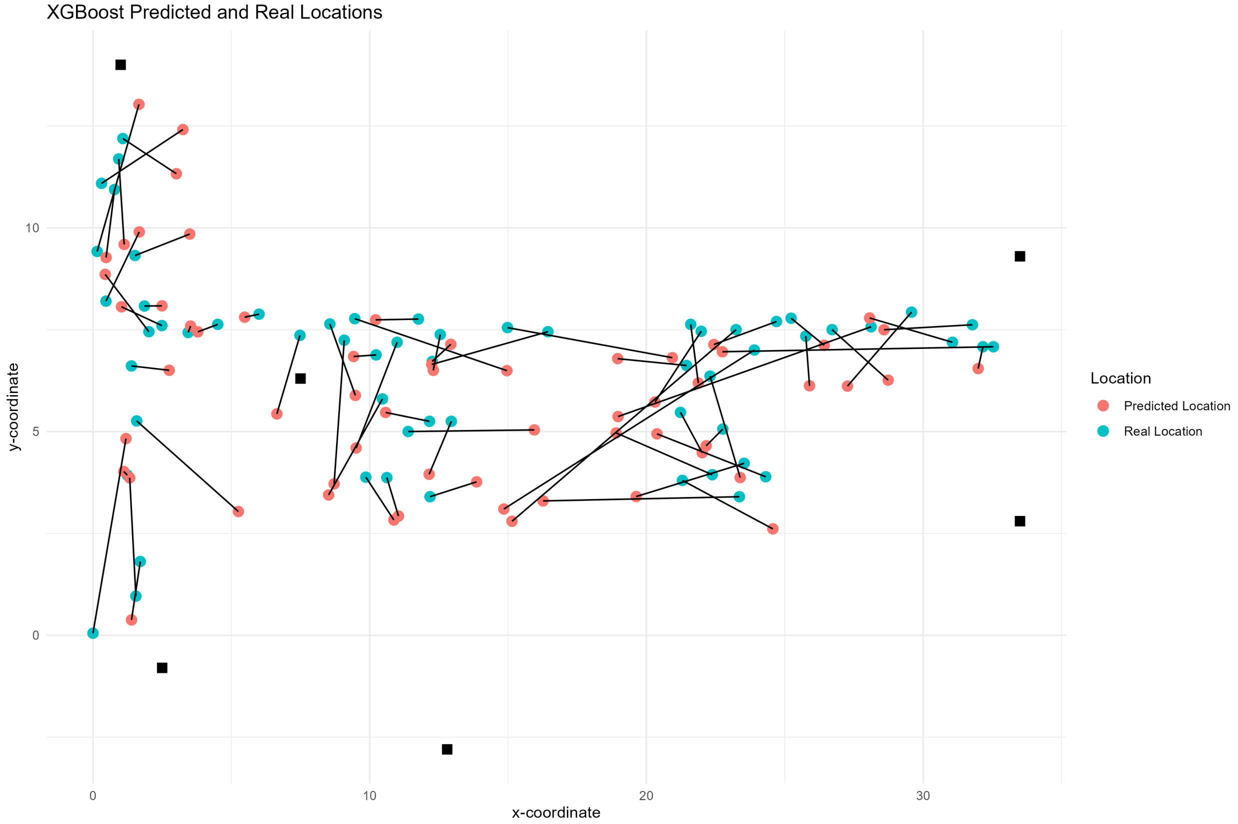

Figure 4: XGBoost Prediction Map. Green dots correspond to real locations for the testing (online) data. Red dots are the predicted locations by XGBoost. Black squares correspond to 6 access points. The error between predictions and real value is represented by black lines.

As expected, KNN is more heavily influenced by the number of adjacent training points. The center of the building is characterized by a higher density of training points in the grid, resulting in the algorithm predicting center values more accurately than those at the far end of the gridded training points. Another influence seems to be the geospatial location of the access points. The west side of the building contains more equally distributed routers. This reduced distance between the testing point and the router results in a stronger strength signal and therefore, allows the algorithm to better triangulate the location. This decrease in distance, and therefore signal strength, also results in potential noise taking over the wanted signal. Similarly, XGBoost is also influenced by this higher-error on the east trend, but to a less extend. The algorithm seems more robust in finding patters when data is sparsely distributed.

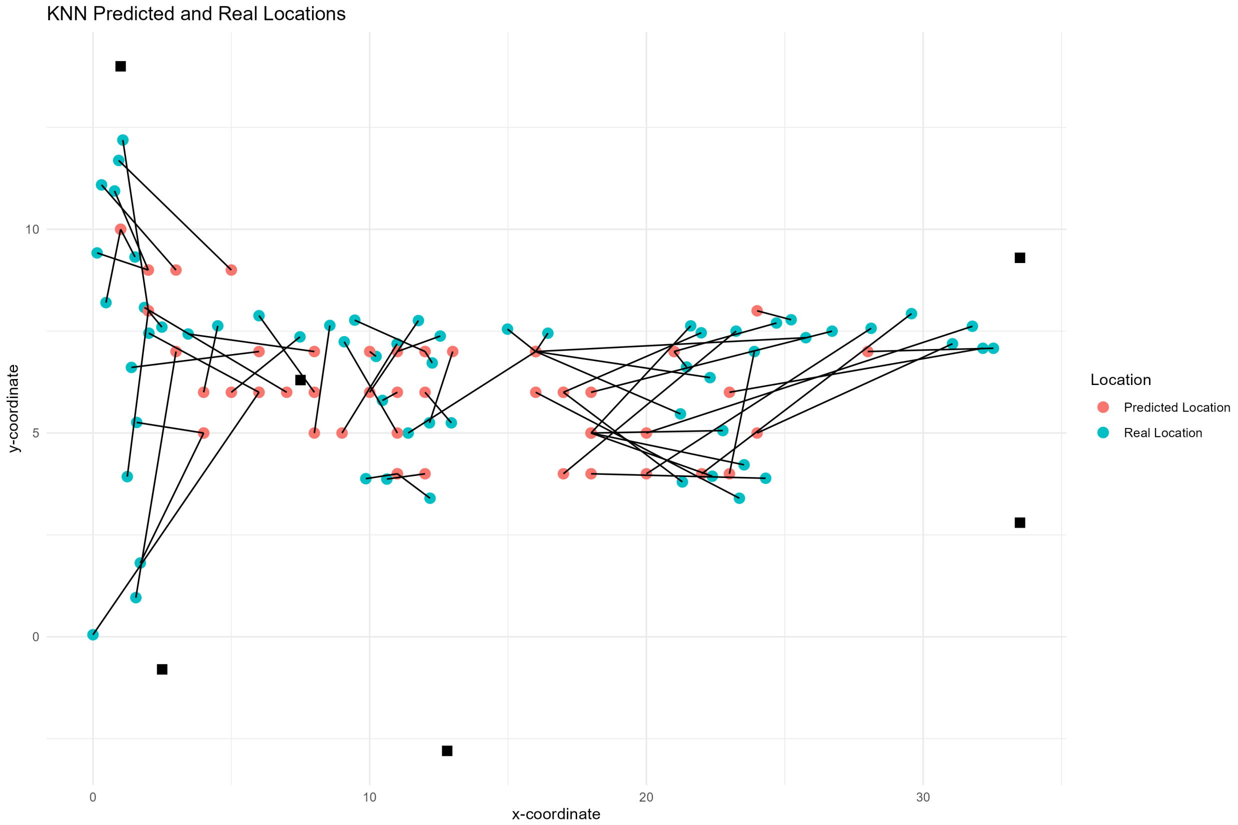

Figure 5: K-Nearest Neighbor Prediction Map. Green dots correspond to real locations for the testing (online) data. Red dots are the predicted locations by XGBoost. Black squares correspond to 6 access points. The error between predictions and real value is represented by black lines.

This issues with high-error predictions might be mitigated by addressing the location of the routers and including a more equally distributed grid. This is a cost-effective way of potentially reducing the error.

6. Conclusions and Recommendations

Both approaches studied offer highly accurate predictions and offer cost-effective alternatives for indoor positioning systems. The wide-spread use of public Wi-Fi signals is an advantage over other signal detection systems, such as Bluetooth. Areas that are closer to access point will be more accurately predicted. If a close monitoring of a device, such as expensive medical equipment, is required, it's suggested that is kept at the proximities of a remote access point. It's important to note the small scale of the study, confined into a 36 by 15 meter area. Further studies are required to understand the robustness of the models in areas with higher room density, and less routers. Another potential source of bias is the use of mean values to create these models, because they might not be representative of the overall signal signature by location. The use of mean values reduced significantly the number of training points, making the models less generalizable.

The general distribution of signal strength for these study included various points that had to be removed due to noise of nearby routers. This excluded points might've included important information that was lost. Better ways of cleaning noisy data and finding valuable patterns in the data set might be expanded to include un-labeled (unsupervised) machine learning.

7. References

[1] Hromadova, Veronika & Machaj, Juraj & Brida, Peter. (2021). Impact of user orientation on indoor localization based on Wi-Fi. Transportation Research Procedia. 55. 882-889. 10.1016/j.trpro.2021.07.056.

[2] Li, You & He, Zhe & Li, Yuqi & Gao, Zhouzheng & Chen, Ruizhi & El-Sheimy, Naser. (2019). Enhanced Wireless Localization Based on Orientation-Compensation Model and Differential Received Signal Strength. IEEE Sensors Journal. PP. 1-1. 10.1109/JSEN.2019.2899895.

[3] P. Bahl and V. N. Padmanabhan, "RADAR: an in-building RF-based user location and tracking system," Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies, Tel Aviv, Israel, 2000, pp. 775-784 vol.2, doi: 10.1109/INFCOM.2000.832252.

[4] Chen, T., and Guestrin, C. (2016). XBGoost: A Scalable Tree Boosting System. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, United States.

Appendices

Appendix A: Individual and General Prediction Errors

The mean average error (MAE) is used to compare the general performance of the two approaches.

Table A1 shows the MAE for XGBoost and KNN on x and y-axis.

A more detailed view of the predicted and true values by cells for XGBoost (Table A2), and KNN (Table 3) demonstrate the high bias in some predicted locations.

Table A1: Error (MAE) on X and Y locations

	KNN	XGB
Error X	3.389667	2.257333
Error Y	1.600833	1.249000

[b]

Table A2: Total Error (MAE)

	KNN	XGB
MAE	2.49525	1.753167

[b]

Table A3: XGBoost model predicted and real locations

X	XGB_pred_X	XGB_pred_Y	XGB_test_X	XGB_test_Y	XGB_Error_X	XGB_Error_Y
0	1.19	4.83	0.00	0.05	1.19	4.78
1	1.66	13.03	0.15	9.42	1.51	3.61
2	3.25	12.41	0.31	11.09	2.94	1.32
3	1.67	9.90	0.47	8.20	1.20	1.70
4	0.48	9.27	0.78	10.94	0.30	1.67
5	1.13	9.59	0.93	11.69	0.20	2.10
6	3.01	11.33	1.08	12.19	1.93	0.86
7	1.11	4.02	1.24	3.93	0.13	0.09
8	2.76	6.50	1.39	6.61	1.37	0.11
9	3.50	9.85	1.52	9.32	1.98	0.53
10	1.33	3.86	1.55	0.96	0.22	2.90
11	5.26	3.04	1.58	5.26	3.68	2.22
12	1.39	0.38	1.71	1.81	0.32	1.43
13	2.49	8.08	1.86	8.08	0.63	0.00
14	9.42	6.84	10.23	6.88	0.81	0.04
15	8.52	3.44	10.46	5.80	1.94	2.36
16	11.04	2.93	10.62	3.87	0.42	0.94
17	9.51	4.60	10.99	7.19	1.48	2.59
18	15.95	5.04	11.39	5.00	4.56	0.04
19	10.21	7.74	11.76	7.76	1.55	0.02
20	10.58	5.47	12.16	5.25	1.58	0.22
21	13.86	3.77	12.18	3.40	1.68	0.37
22	12.93	7.14	12.26	6.72	0.67	0.42
23	12.30	6.50	12.55	7.38	0.25	0.88
24	12.15	3.95	12.95	5.25	0.80	1.30
25	20.94	6.81	14.98	7.55	5.96	0.74
26	12.24	6.65	16.44	7.45	4.20	0.80
27	0.44	8.86	2.02	7.45	1.58	1.41
28	1.03	8.06	2.49	7.60	1.46	0.46
29	22.02	4.48	21.23	5.47	0.79	0.99
30	24.57	2.61	21.30	3.80	3.27	1.19
31	18.96	6.79	21.45	6.62	2.49	0.17
32	21.87	6.19	21.60	7.63	0.27	1.44
33	20.31	5.72	21.98	7.46	1.67	1.74
34	23.39	3.87	22.30	6.36	1.09	2.49
35	18.91	4.97	22.38	3.94	3.47	1.03
36	22.16	4.65	22.76	5.06	0.60	0.41
37	15.14	2.80	23.24	7.50	8.10	4.70
38	16.26	3.30	23.36	3.40	7.10	0.10
39	19.63	3.40	23.53	4.22	3.90	0.82
40	14.85	3.10	23.90	7.00	9.05	3.90
41	20.38	4.94	24.31	3.89	3.93	1.05
42	22.44	7.14	24.70	7.70	2.26	0.56
43	26.43	7.12	25.23	7.78	1.20	0.66
44	25.89	6.12	25.76	7.34	0.13	1.22
45	28.74	6.26	26.71	7.50	2.03	1.24
46	18.98	5.37	28.12	7.57	9.14	2.20
47	27.27	6.12	29.58	7.93	2.31	1.81
48	3.53	7.59	3.44	7.43	0.09	0.16
49	28.07	7.79	31.06	7.19	2.99	0.60
50	28.59	7.50	31.78	7.62	3.19	0.12
51	31.99	6.55	32.16	7.08	0.17	0.53
52	22.75	6.96	32.54	7.08	9.79	0.12
53	3.79	7.45	4.51	7.63	0.72	0.18
54	5.48	7.81	6.00	7.88	0.52	0.07
55	6.65	5.43	7.48	7.36	0.83	1.93
56	9.48	5.89	8.56	7.64	0.92	1.75
57	8.71	3.72	9.08	7.24	0.37	3.52
58	14.96	6.49	9.46	7.77	5.50	1.28
59	10.87	2.83	9.86	3.88	1.01	1.05

[b]

Table A4: KNN model predicted and real locations

X	Knn_pred_X	Knn_pred_Y	Knn_test_X	Knn_test_Y	Knn_Error_X	Knn_Error_Y
0	6	6	0.00	0.05	6.00	5.95
1	2	9	0.15	9.42	1.85	0.42
2	3	9	0.31	11.09	2.69	2.09
3	1	10	0.47	8.20	0.53	1.80
4	2	9	0.78	10.94	1.22	1.94
5	5	9	0.93	11.69	4.07	2.69
6	2	8	1.08	12.19	0.92	4.19
7	2	8	1.24	3.93	0.76	4.07
8	6	7	1.39	6.61	4.61	0.39
9	1	10	1.52	9.32	0.52	0.68
10	3	7	1.55	0.96	1.45	6.04
11	4	5	1.58	5.26	2.42	0.26
12	4	5	1.71	1.81	2.29	3.19
13	7	6	1.86	8.08	5.14	2.08
14	10	7	10.23	6.88	0.23	0.12
15	11	6	10.46	5.80	0.54	0.20
16	12	4	10.62	3.87	1.38	0.13
17	9	5	10.99	7.19	1.99	2.19
18	16	7	11.39	5.00	4.61	2.00
19	10	6	11.76	7.76	1.76	1.76
20	13	7	12.16	5.25	0.84	1.75
21	11	4	12.18	3.40	1.18	0.60
22	12	7	12.26	6.72	0.26	0.28
23	11	7	12.55	7.38	1.55	0.38
24	12	6	12.95	5.25	0.95	0.75
25	16	7	14.98	7.55	1.02	0.55
26	16	7	16.44	7.45	0.44	0.45
27	6	6	2.02	7.45	3.98	1.45
28	2	8	2.49	7.60	0.49	0.40
29	16	7	21.23	5.47	5.23	1.53
30	17	6	21.30	3.80	4.30	2.20
31	21	7	21.45	6.62	0.45	0.38
32	18	5	21.60	7.63	3.60	2.63
33	17	6	21.98	7.46	4.98	1.46
34	16	7	22.30	6.36	6.30	0.64
35	18	5	22.38	3.94	4.38	1.06
36	18	5	22.76	5.06	4.76	0.06
37	17	4	23.24	7.50	6.24	3.50
38	16	6	23.36	3.40	7.36	2.60
39	18	5	23.53	4.22	5.53	0.78
40	23	4	23.90	7.00	0.90	3.00
41	18	4	24.31	3.89	6.31	0.11
42	21	7	24.70	7.70	3.70	0.70
43	24	8	25.23	7.78	1.23	0.22
44	16	7	25.76	7.34	9.76	0.34
45	18	6	26.71	7.50	8.71	1.50
46	20	4	28.12	7.57	8.12	3.57
47	22	4	29.58	7.93	7.58	3.93
48	8	7	3.44	7.43	4.56	0.43
49	24	5	31.06	7.19	7.06	2.19
50	20	5	31.78	7.62	11.78	2.62
51	23	6	32.16	7.08	9.16	1.08
52	28	7	32.54	7.08	4.54	0.08
53	4	6	4.51	7.63	0.51	1.63
54	8	6	6.00	7.88	2.00	1.88
55	5	6	7.48	7.36	2.48	1.36
56	8	5	8.56	7.64	0.56	2.64
57	11	5	9.08	7.24	1.92	2.24
58	12	7	9.46	7.77	2.54	0.77
59	11	4	9.86	3.88	1.14	0.12