Indoor Positioning Based on Wireless Fingerprinting

18416144 HUANG Jiawen

18405606 KE Junxing

Abstract

With the popularization of large indoor places in people's lives, the positioning service technology applied indoors is growing. Unlike outdoors that use GPS technology for positioning, walls of the building make indoor GPS positioning extremely inaccurate. In order to provide accurate positioning indoors, the article uses wireless fingerprinting technology, concentrating on obtaining accurate indoor positioning models through training data sets. In our work, we select indoor spaces with typical characteristics, and divide the indoor space into equal proportions of small areas represented by fingerprints based on the signal strength of each wireless router. Then, we use SVM, KNN and decision tree algorithms to train the data we get. Finally, through analysis and comparison, the optimal model can be obtained.

1. Introduction

I. Background

Nowadays, many applications provide personalized services based on the user's location. The positioning function of most applications is implemented by GPS technology. However, GPS technology does not provide accurate positioning indoors because of the large number of walls and steels. Thus, in order to provide users with accurate indoor positioning, many indoor positioning technologies have been developed. Wireless fingerprinting is one of the technologies. The advantage of this technology is that no additional hardware support is required for the reason that the technology implement by the received signal strengths and now many indoor venues are equipped with many wireless signals. But the wireless fingerprinting needs to collect vast wireless signal data indoors to train the model, which is very time consuming. A highly accurate model is needed to take advantage and avoid shortcomings of this technology.

II. Exist Methods

Dan Li, Le Wang, et al have explored the indoor positioning system design and used different algorithms to improve the accuracy of indoor positioning. According to their research, KNN shows extraordinary performance when the sample data set is larger than 1000. The Gradient Boosting algorithm has subtle cross-validation errors when dealing with small amounts of data, but is robust to missing data[1]. Another solution is presented by Beomju Shin, et al. Instead of applying kNN algorithm by giving a fixed number k of the neighbors, they change the considered neighbors number and assign them weights[2]. Jan

Racko, Juraj Machaj, et al have advocated the application of the interpolation to reduce the time needed for indoor positioning. Smaller areas indoor which represented by fingerprints will be calculated by linear interpolation algorithm and Delaunay algorithm[3].

III. Novelty

The traditional method of fingerprint required to collect vast data to train the model, which is a very heavy workload. Therefore, we want to achieve better balance between the accuracy and workload.

Inspired by previous work, we have improved the weighting method of weighted KNN. Experiments show that the algorithm can effectively reduce the error, and the problem can get better performance than SVM and decision tree.

2. Methodology

I. Data Collection and Data Processing

In this project, we decided to collect data and process the data ourselves instead of using public data set. The experiment was conducted on the 2nd floor of WLB, Hong Kong Baptist University. The size of space is 112m*44m, as shown in figure 1. We use the smartphone produced by SONY, the operating system is installed with Android8.0. We go to each reference point to collect the Received Signal Strength Index (RSSI) of each Access Point (AP) that can be sensed by the device. These RSSI data and location information as labels are then stored in the Realm database.

The first thing we need to do is to collection data of RSSI in order to train a prediction model in this part. Usually, the data will be split into two section according to specific proportion after collection, training data set, and testing data set respectively. However, we are considering it from the beginning when we are doing this project. First, we map the map of experiment site to a two-dimensional space axis. The training set data is set at one interval of 4 meters in two axial directions, and the test set data is collected at intervals of 6 meters, as shown in figure 2. Other conditions are the same as the training set.

After that we have to process the data, and the collected data is formatted for the next training and testing. Because the access points around different reference points are different, they can be distinguished by Mac address. However, all reference points should have the same features, so APs that are not received at different reference points are filled with -100 values, as shown in figure 3.

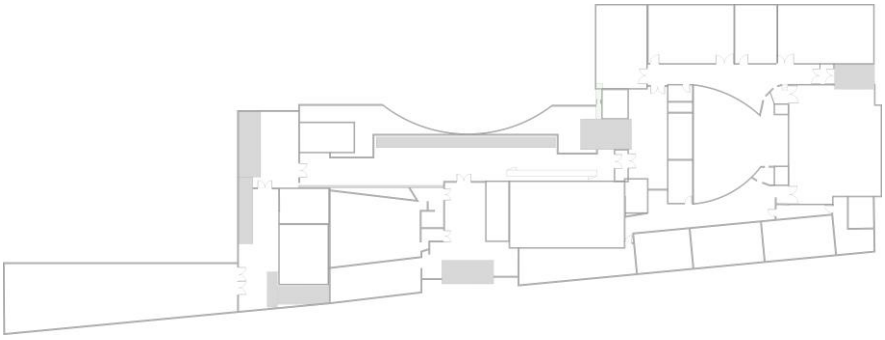


Figure 1: Floor plan of WLB 2nd floor

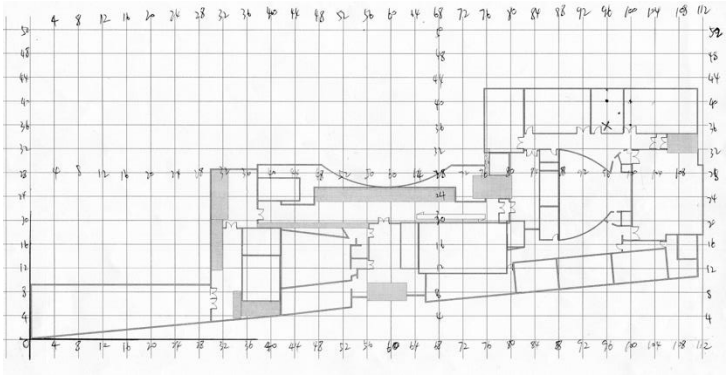
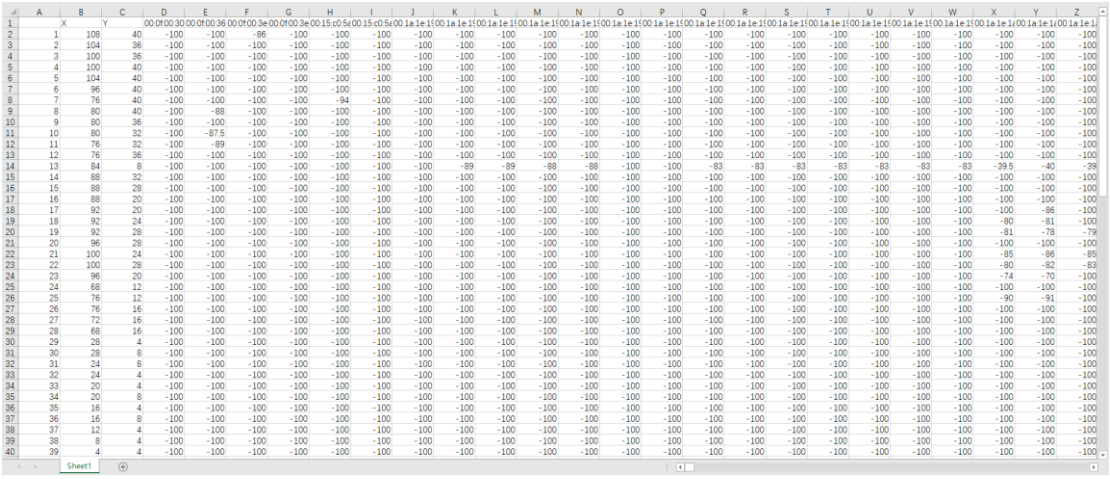


Figure 2: Floor plan with coordinate system



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1		X	Y	00:0f:00:30	00:0f:00:36	00:0f:00:3e	00:0f:00:3e	00:15:c5:5e	00:15:c5:5e	00:1a:1e:1f	00:1a:1e:1f	00:1a:1e:1f	00:1a:1e:1f	00:1a:1e:1f	00:1a:1e:1f	00:1a:1e:1f	00:1a:1e:1f	00:1a:1e:1f	00:1a:1e:1f	00:1a:1e:1f	00:1a:1e:1f	00:1a:1e:1f	00:1a:1e:1f	00:1a:1e:1f	00:1a:1e:1f	00:1a:1e:1f	00:1a:1e:1f
2	1	108	40	-100	-100	-86	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
3	2	104	36	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
4	3	100	36	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
5	4	100	40	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
6	5	104	40	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
7	6	96	40	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
8	7	76	40	-100	-100	-100	-100	-94	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
9	8	80	40	-100	-88	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
10	9	80	36	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
11	10	80	32	-100	-87.5	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
12	11	76	32	-100	-89	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
13	12	76	36	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
14	13	84	8	-100	-100	-100	-100	-100	-100	-100	-89	-89	-88	-88	-100	-100	-83	-83	-83	-83	-83	-83	-83	-39.5	-40	-39
15	14	88	32	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
16	15	88	28	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
17	16	88	20	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
18	17	92	20	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-86	-100
19	18	92	24	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-80	-81	-100
20	19	92	28	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-81	-78	-79
21	20	96	28	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
22	21	100	24	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-85	-86	-85
23	22	100	28	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-80	-82	-83
24	23	96	20	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-74	-70	-100
25	24	68	12	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
26	25	76	12	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-90	-91	-100
27	26	76	16	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
28	27	72	16	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
29	28	68	16	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
30	29	28	4	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
31	30	28	8	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
32	31	24	8	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
33	32	24	4	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
34	33	20	4	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
35	34	20	8	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
36	35	16	4	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
37	36	16	8	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
38	37	12	4	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
39	38	8	4	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100
40	39	4	4	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100

Figure 3: Training data set

II. KNN

K-nearest neighbor algorithm is a basic classification and regression method. We mainly apply the classification of it. Given a set of wireless signal data, for the newly input wireless signal data, we find k number signals that are closest to the newly input signal in the dataset. If most of these k number signals belong to a certain class, we will classify the newly input signal data into the class. We use the Euclidean distance as a measure of distance.

Weighted KNN is one of candidate algorithm, which is suitable for this topic and has potential to achieve high performance. In order to make the reference point with high similarity contribute a lot to the prediction point, we have made some improvements to the weight KNN to better adapt to our data. Different from the weighting method of the literature[4], it is as follows. The TD is the sum of the distances from all reference points to the predicted points, and d_i is the distance from the reference point to the predicted point, using the Euclidean distance. Besides, α is a parameter, greater than 0.

$$TD = \sum_{i=1}^{k} (\frac{1}{d_i})^{\alpha}, \ \ \alpha > 0 \qquad (1)$$

$$\text{Similarity}_i = \frac{(\frac{1}{d_i})^{\alpha}}{TD}, \alpha > 0 \qquad (2)$$

There are four advantages of KNN algorithm in this topic:

1. Easy to use and understand, one of the simplest algorithms in machine learning.
2. Can be used for both numeric and discrete data
3. Excellent performance on multi-classification issues.
4. Training time complexity is O(n)

However, the KNN algorithm has disadvantages, here is it.

1. Lazy learning algorithm, high memory overhead
2. Sample value imbalance problem. Feature normalization is required, adding additional computational effort.
3. Poor interpretability.

III. SVM

Support Vector Machine is a supervised learning model related to machine learning. It can be used for classification and regression analysis. We mainly apply the SVM classification method, combined with the kernel function to nonlinearly classify the collected wireless signal data, and finally trains to obtain a classification prediction model that input the wireless signal data and output the class of the location.

Advantages:

1. It is very efficient in high dimensional space.
2. The algorithm is still valid when the data dimension is larger than the number of samples.
3. Efficient use of memory.

4. Versatility: Different kernel functions correspond to specific decision functions. People can use either the common kernel already provided, or the custom kernel.

Disadvantages:

1. Need to avoid overfitting when choosing a kernel function.
2. It does not directly provide probability estimates and requires additional cross-validation calculations to derive estimates.

IV. Decision Tree

The decision tree is a tree structure. Each non-leaf node of the tree represents a filter condition on a feature attribute, each branch represents the output of the feature attribute on a range of values, and each leaf node represents a class. For the newly input wireless signal data, when passing through each non-leaf node, the class of data is judged according to the condition in the non-leaf node.

Advantages:

1. The decision tree is highly explanatory and easy to understand and use.
2. Data for decision tree learning generally does not require preprocessing.
3. Can be used for both numeric and discrete data
4. The decision tree is a white-box model that makes it easy to derive logical expressions for the model.
5. The efficiency is high, and the decision tree only needs to be built once and used repeatedly

Disadvantages:

1. Sample value imbalance problem. The result of information gain is biased towards those features with more values.
2. It is easy to cause overfitting.
3. Difficult to process data with strong feature relevance.
4. Less accurate than other machine learning methods.

3. Experimental Study

In this project, we use the same criteria to evaluate the performance of the algorithm, that is, to compare the error between the predicted value and the actual value. In the test, we introduced 3 dimensions, including minimum error, maximum error and average error.

I. KNN

In the KNN algorithm, we obtain the optimal solution by adjusting the K value and the α value. According to the results of our many experiments, the effect of K value on the minimum error is obvious, as shown in figure 4. From Figure 4, we can see that the average error is the smallest when k is taken as 3.

When the optimal k value is found, the next step is to determine the value of α. The experimental results are shown in Figure 5. The maximum error value tends to decrease as α increases. However, the minimum error shows a different change, first decreasing and then increasing. The average error also shows a trend similar to that of the minimum error. Combining these three dimensions, we can see from Figure 5 that when α is 4, the error is the smallest.

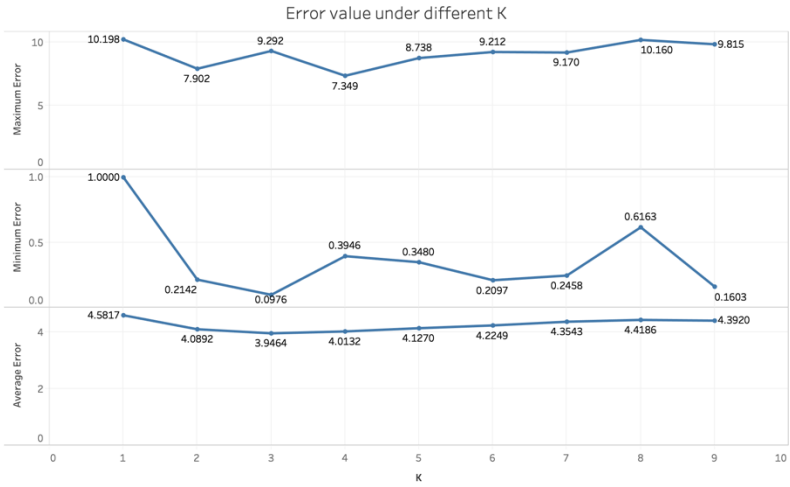


Figure 4: Errors at different K values

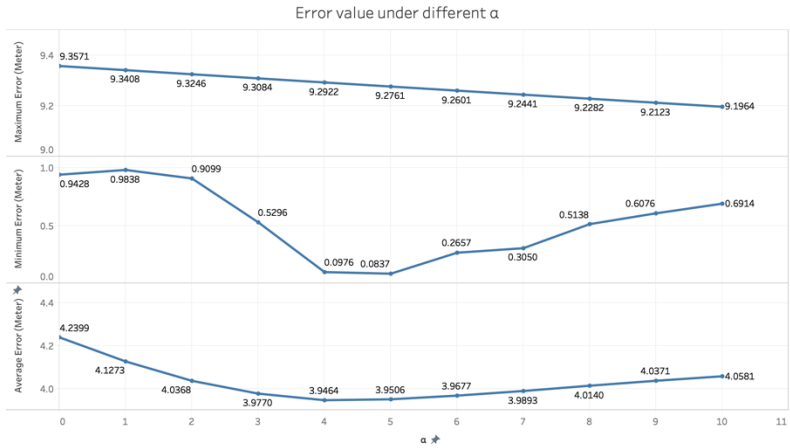


Figure 5: Errors at different α values

II. SVM

We selected 4 different kernels for testing, including linear, poly, rbf and sigmoid. The results obtained are shown in Table 1. From Table 1, we can see that the linear, ploy and rbf

kernels perform almost the same in terms of maximum error and minimum error. In addition, the average error of the rbf and linear kernels also showed the same result. Besides, the average error calculated by the sigmoid kernel is as high as 30 meters, indicating that the algorithm is not suitable for this problem.

Kernel	Minimum Error	Maximum Error	Average Error
linear	1	10.19803903	4.581663905
poly	1	10.19803903	4.704481045
rbf	1	10.19803903	4.581663905
sigmoid	2	60.8276253	30.62751196

Table 1: SVM test results

III. Decision Tree

We chose two different branching algorithms of the decision tree and tested the performance under different criterion. The results are shown in table 2. The minimum error is also more than 7m, which is much larger than the SVM's 4.7m error.

Decision Trees	Minimum Error	Maximum Erro	Average Error
Classifier-gini	2	113.2254388	30.16108417
Classifier-entropy	2	84.02380615	16.88344713
Regression-mse	1	84.09518417	12.5629438
Regression-friedman_mse	1.2	84.09518417	12.20303288
Regression-mae	2	27.78488798	7.28762355

Table 2: Decision tree test results

As we can see from table 3, weighted KNN performs the best in this scenario, followed by SVM, and the worst is decision tree.

Methods	Minimun Error(Meter)	Maximum Error (Meter)	Average Error
Weighted KNN	0.097585087	9.292226528	3.946371018
SVM	1	10.19803903	4.581663905
Decision Tree	2	27.78488798	7.28762355

Table 3: Comparison of test results of three algorithms

4. Conclusion

According to the comparison of the accuracy of models trained by three different machine learning algorithms, the KNN algorithm perform the best accuracy among the three algorithms, given the space of the 2nd floor of WLB and reasonable data space. Moreover, the error bias of the kNN is much smaller than the other two algorithms. Although not good enough for kNN, the SVM perform good when using linear, poly and rbf kernel. Decision tree has a poor performance to be a model for predicting the location and its maximum error

can reach an incredible 27 meters. As the conclusion we get, in order to acquire high accuracy of the indoor positioning model, we propose the weighted KNN algorithm for the indoor positioning.

Reference

[1] Dan Li, Le Wang, Shiqi Wu, “Indoor Positioning System Using Wifi Fingerprint”, *Stanford University*.

[2] B Shin, J H Lee, T Lee et al., "Enhanced weighted K-nearest neighbor algorithm for indoor Wi-Fi positioning systems", *Proc. the 8th International Conference on Computing Technology and Information Management,* pp. 574-577, April 2012.

[3] J. Racko, J. Machaj, P. Brida, “Wi-fi fingerprint radio map creation by using interpolation”, Procedia Eng 192, 753–758 (2017)

[4] B. S., -, J. H. L., -, T. L., & -, H. S. K. (2012). Enhanced Weighted K-Nearest Neighbor Algorithm for Indoor Wi-Fi Positioning Systems. *International Journal of Networked Computing and Advanced Information Management*, *2*(2), 15–21. https://doi.org/10.4156/ijncm.vol2.issue2.2