A Site Selection and Recruitment Strategy of JPTA Program Evaluation: Final Report

DS4SI Assignment 2

Tong Jin
NYU Steinhardt A3SR

September 30, 2019

Overview

This study discusses a site selection strategy of the JPTA program and its design process. It covers details about principles, priorities, statistical computation, and simulation.

The strategy is to calculate the total number of sites that need to be reached in order to successfully recruit 30 sites. It also determines which site should we reach. Specifically, the strategy is based on the following principles

1. The sites selected should be well balanced to represent a reasonable counterfactual.

2. The selected sites should have enough comfort levels to ensure that the recruitment goal can be achieved.

3. The comfort level of selected sites should be as high as possible.

4. The total cost should be as low as possible.

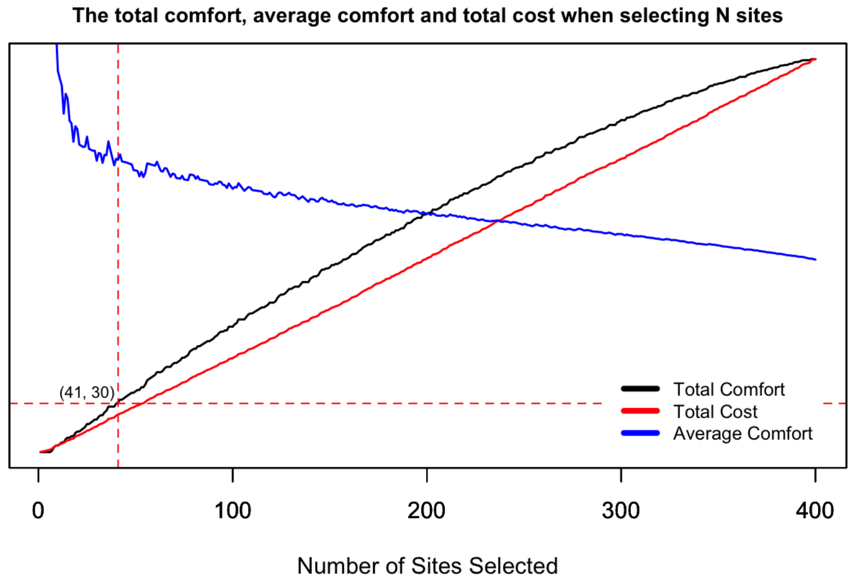
Based on these principles, the strategy determines several priorities. The top priority is the balance of selected sites. This is because a representative sample ensures the validity of the study. With a balanced sample, we can evaluate the program outcomes with persuasive results. The second priority is the comfort level. Since the probability that a site accepts the evaluation is equal to the index of comfort level, we assume that the higher the comfort level, the more likely the site can be successful recruited. Thus, we need to focus on those sites with high comfort levels to ensure a high quality recruitment. Moreover, the cost control is another priority. Keeping the cost at a reasonable level can increase the flexibility of follow ups and possible strategy alternations. Therefore, this study discusses a selection strategy with balanced sample distribution, highest comfort level and lowest total cost.

Site Selection Strategy Details

First of all, the provided sites in the dateset are randomly selected by the JPTA program. To achieve a balanced sample, the strategy focuses on the *region* and *urban* variables. The first step is to divide the JPTA dateset into 4 subsets grouped by regions. This is to ensure the geographic balance. Second, each regional set is divided into 2 subsets grouped by the *urban* variable. This is to ensure the balance between urban and rural areas. Then, all sets are sorted by the comfort level in a descending order. Sites with higher comfort levels are listed first.

The second step is to determine the weight of each subset. The weights of subsets are apportioned among the subsets by site population. They are calculated by the site populations divided by the total population. With weights determined, the study implements the next step through a simulation of different number of sites selected. It calculates how many site samples with highest comfort level in each set should be selected under a certain number of total sites selected. Then, the study pulls out these sample sites and calculates the total comfort index, the average comfort level, and the total cost.

Next, the study plots the total comfort index, the total cost, and the average comfort level as curved lines and combines them into a single graph, as follows.



From the graph, the study interprets that there is a positive relationship between total comfort level and number of sites. The total comfort level increases when total number of sites selected increases. However, the slope is decreasing, meaning that the amount of increase in total comfort for each increase in number of sites is decreasing. This also means that we should focus on the first half. Moreover, because the comfort level represents the success rate of site recruitment, the rounded sum of comfort levels is equal to the number of sites that can be successfully recruited. When the number of sites selected equals to 41, the total comfort level reaches 30, which is the recruitment goal. Therefore, we need to at least recruit 41 sites in order to meet our recruitment goal.

Another interpretation is that the average comfort level decreases when the total number of sites increases. At a range from 41 to 400, when the total number of sites equals to 42, the average comfort level reaches the highest point. When the total number of sites equals to 41, the average comfort level reaches the second highest point.

Furthermore, The total cost increases as the total number of sites increases. At a range from 41 to 400, when the total number of sites equals to 41, the total cost reaches the lowest point.

Based on these interpretations, the study determines that when the number of sites is 41, the selected sites reach the best combination such that:

a. The total comfort level is greater than 30.

b. The total cost is the lowest.

c. The average comfort level is higher than most of other combinations.

Thus, **the expected number of sites is 41. The total cost is $84337.63.**

Samples

The following is a list of 41 samples selected from the ‘JPTA‘ dateset through the strategy discussed above.

No.	Site ID	Region	Urban	Distance	Comfort
1	1085	4	0	2.553477201	0.679350879
2	1374	2	0	2.271912984	0.787117751
3	1802	3	0	2.794900244	0.790650855
4	1895	2	0	2.724818682	0.682376486
5	2082	4	0	2.070033644	0.694442528

No.	Site ID	Region	Urban	Distance	Comfort
6	2096	4	1	0.982813371	0.750531719
7	2708	1	1	0.836943262	0.714475664
8	2816	1	0	1.38429861	0.765441226
9	2901	3	0	2.787163433	0.767268036
10	2951	1	0	1.813174836	0.712335428
11	3075	2	1	0.623901741	0.679342539
12	3232	1	0	2.100472634	0.932860125
13	3396	4	1	1.086396809	0.742787254
14	3860	4	0	1.853656549	0.712177717
15	4118	1	0	1.980030161	0.668846621
16	4429	3	0	2.825837536	0.821131199
17	4540	1	1	0.802943183	0.76018279
18	4807	4	0	1.943289476	0.754285701
19	5091	4	0	2.730132873	0.677503654
20	5095	4	0	2.385872745	0.704596853
21	5263	1	0	1.536252672	0.726100397
22	5336	4	0	1.99071893	0.780115529
23	5454	3	1	2.339106278	0.736161544
24	5514	2	0	2.245420459	0.808873627
25	5545	2	0	2.204668964	0.660641277
26	5767	1	0	1.957446084	0.69420575
27	5907	2	0	1.449535404	0.649801936
28	5948	3	0	2.860670869	0.711959161
29	5977	1	0	2.392631301	0.705663862
30	6533	3	0	3.743549048	0.837499189
31	6759	1	0	1.909372356	0.744047603
32	7154	4	0	1.972726988	0.919764796
33	7440	3	0	3.388618562	0.819853889
34	8164	3	0	3.123416364	0.94020332
35	8648	1	0	2.479525599	0.685968815
36	8744	4	0	2.005962568	0.715452428
37	9446	2	0	1.775602529	0.745056418
38	9582	3	1	1.865594907	0.717615916
39	9670	3	0	2.961339567	0.722016288
40	9681	2	0	2.05486302	0.671470923
41	9714	4	0	1.866163364	0.731437269

Discussion

To conclude, the study applies a site selection strategy to determine that 41 sites representing different regions and both urban and rural areas should be selected from the ‘JPTA‘ dateset. The total cost of this strategy is the lowest among all other choices.

There are several limitations in this study. First, the study primarily focuses on geographic balancing and excludes the potential impacts of education level, income level, unemployment rate, and availability of other programs. Although the influences of these variables are weakened by JPTA’s randomization, the externality of excluding these variables cannot be ignored.

Second, the study performs a trade-off between cost efficiency and representativeness. To achieve the best level of representativeness, sites should be randomly selected in different regions and areas. However, this would greatly increase the cost because those sites with significant low comfort indexes would probably refuse to participate, resulting the waste of money and human resources. Therefore, to address concerns regarding cost efficiency, the study adopts a strategy to select sites with top comfort indexes.

Appendix

A Site Selection and Recruitment Strategy of JPTA Program Evaluation: Data Analysis.pdf