taipei_passerby_prediction

October 11, 2020

1 Taipei Passerby Prediction

We would like to know the market size / demand. With that in mind, we would use the number of passerby as one of the factor (later would be combined with the people buying power). To estimate that, the data we would use is:

- **Taipei MRT passenger data from gov public data site.** 2 years of Taipei MRT data that show how many passenger come and go in MRT station on hourly basis.

The way we derive passerby from the data are:

- **Try represent MRT for starting activities only.** We will limit the scope of the data to early hour only.
- **Simulate walking distance on MRT passenger.** We would try to heuristically simulate the movement of the people there
- **Aggregate the passerby on village level detail.** We would aggregate the data into village level for analysis usage.

1.1 Use only specific hours to estimate passerby number

To get the passenger data, we would use the traffic which accounted for the people that would start their activities only. So we would exclude the number of people that would use the MRT to end their day (i.e. go home, etc). Therefore we would determine the hour we want to take on.

We would check the hour based on people activities, some of our consideration include: - **Average peak hour for going to work**, we don't want to set the hour befire this time.

To make and confirm out hypothesis, we would do - **Get hourly average number of people that use MRT per hour.** We would split based on weekdays and weekend

1.1.1 Get hourly average number

Aggregate hourly MRT passenger, split by weekdays and weekend.

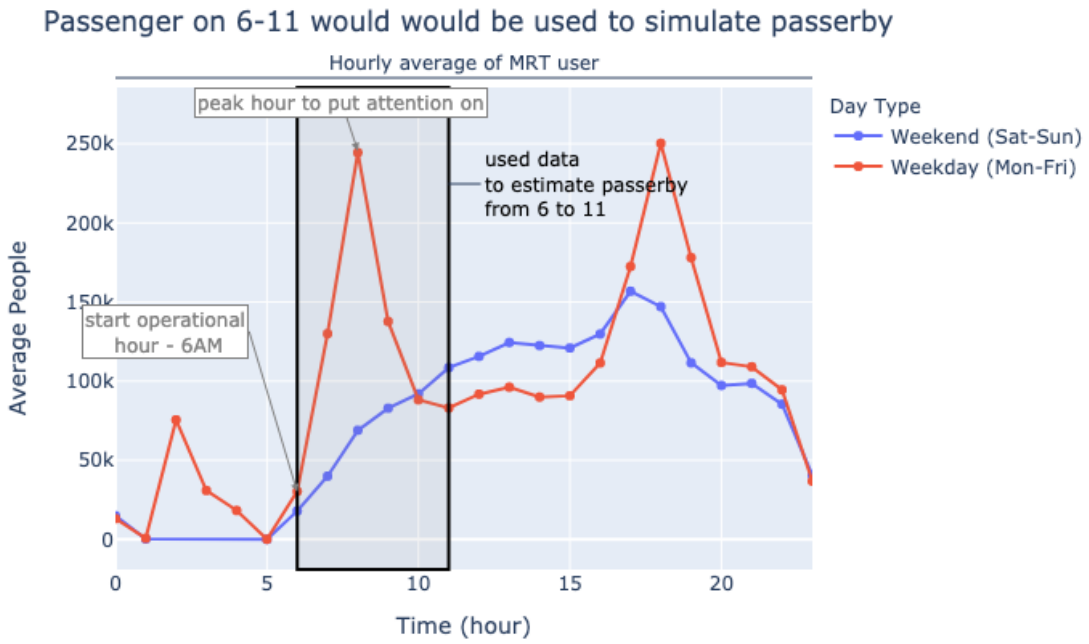
here is the example of the table:

	mon_to_ friday	time_period	person_times
0	0	0	14948

```
|  1 |         0 |             1 |       348.93   |
|  2 |         0 |             5 |        63.0412 |
|  3 |         0 |             6 |     17964.5    |
|  4 |         0 |             7 |     39963.9    |
+----+-----------+---------------+----------------+
```

1.1.2 Data visualization

Here is how the data looks like.



1.1.3 Conclusion

To estimate passerby number or other stuff, we would use MRT data only in **6 until 11**

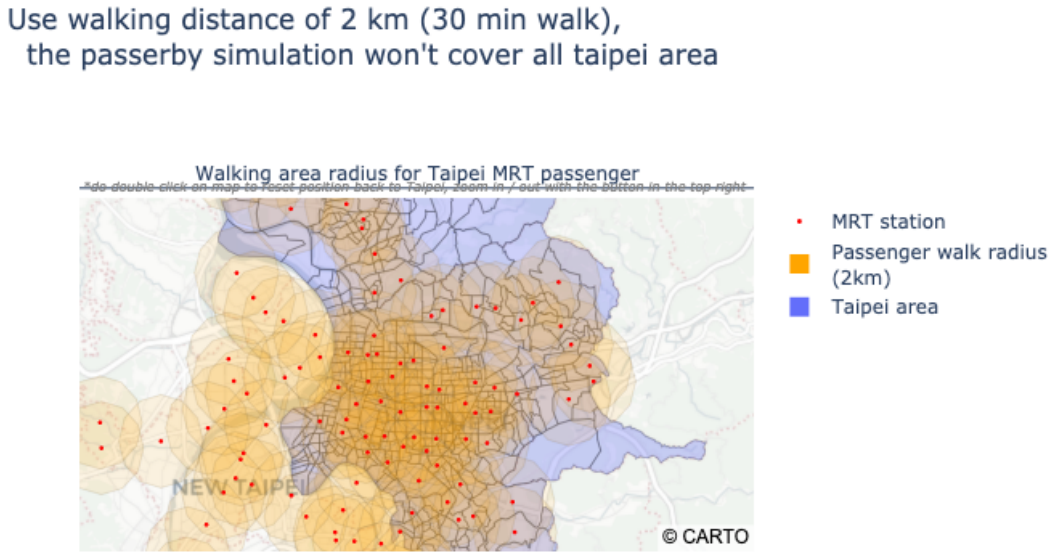
1.2 Set radius on passenger walk distance

As people would also move around, therefore we try to simulate their movement. In this analysis we heuristically try to simulate their movement with:

- **Set walking radius.**
- **Create a probability distribution formula.**

1.2.1 Data visualization

Here is the maximum radius of 30 min walks (2 km) looks like.



1.2.2 Distribution Formula

We would use the 2 km radius, which still possible (30 mins walks), even it is limited to some area. But we would create a really steep distribution function.

The distribution formula is based on:

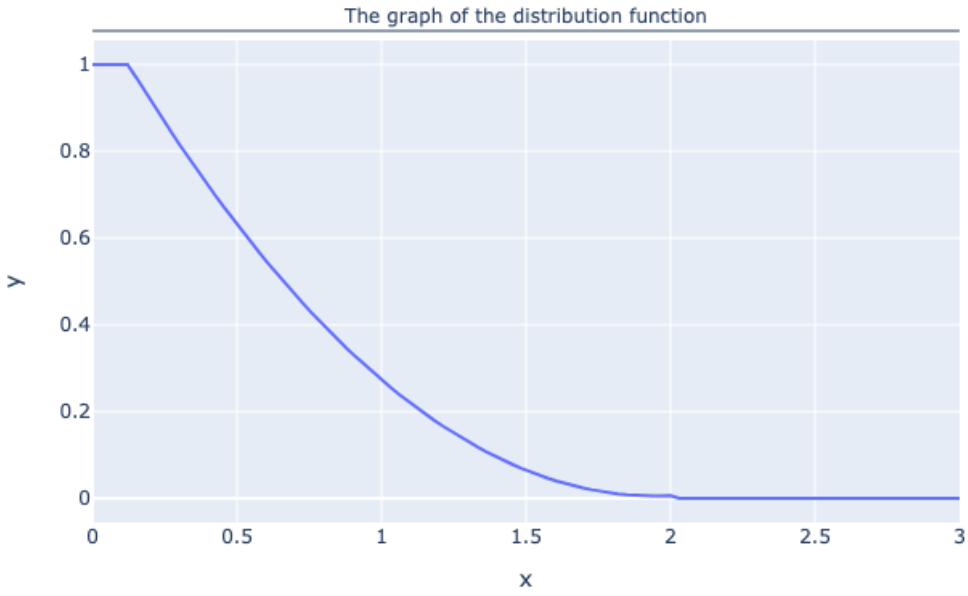
- previous maximum determined radius, 2 km
- use quadratic from about this heuristic logic $f(0) = 1$, $f(2) = 0$, $f(0.3) = 0.7$, $f(1) = 0.2$ that limit x from 0-2 with minimum value of 0 and maximum value of 1:

Therefore, the distribution value that we get is:

$$f(x) = \frac{3x^2}{10} - \frac{17x}{6} + 1 \tag{1}$$

with result no more than 1 or less than 0

The probability function would less likely to pass 1 km. Here is how the distribution looks like.



1.3 Predicting population passerby number

From the previous analysis, we would need to translate the sample data into the real / population data. Therefore we would do:

- **Get population / sample data ratio.** We would make a sample-to-population multiplier for it.
- **Put data into walking simulation formula.** We fit the data into the previous movement formula.

1.3.1 Population / sample data ratio

We use this fact:

- Goverment data in show that in 2019, the public transportation usage in Taipei was 49.4%
- Goverment data show that Taipei population in 2016 was 2,695,704 people
- Goverment data show that Taipei aging population (above 65) in 2016 was 419,130

Therefore, we would use the current sample (taipei MRT data) to predict the population using the fact. We would use this assumption:

- Current MRT data distribution represent at least 80% overall Taipei people activities
- Ignore the 20% unrepresented population passerby
- Aging population would not having any activities at all (use for reducing overall passerby number)

$$sample\text{-}to\text{-}population\ multiplier = \frac{(Taipei\ population * 0.8) - Taipei\ elderly\ population}{Total\ sample\ data} \tag{2}$$

(3)

The calculation on cell below would show the result is 2.75. Therefore we would use the multiplier 2.75 to predict passerby (population) data.

1.3.2 Movement / walking simulation

Based on previous part, we would use this distribution formula: the distribution value that we get is:

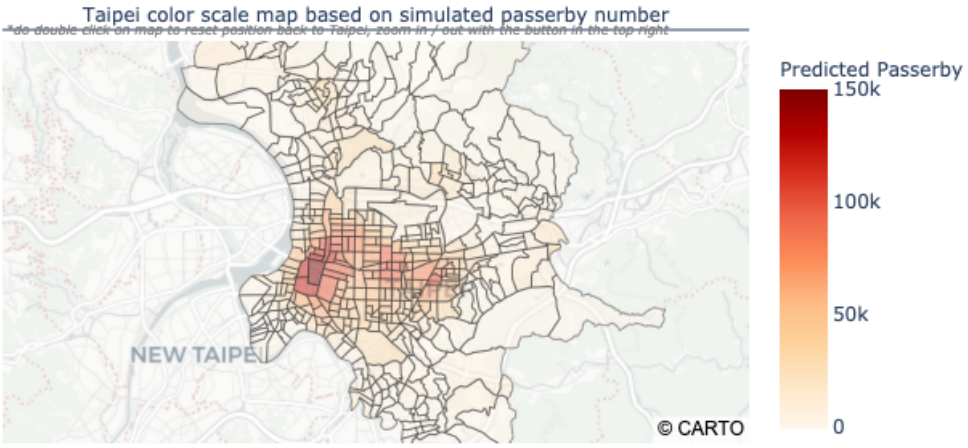
$$f(x) = \frac{3x^2}{10} - \frac{17x}{6} + 1 \tag{4}$$

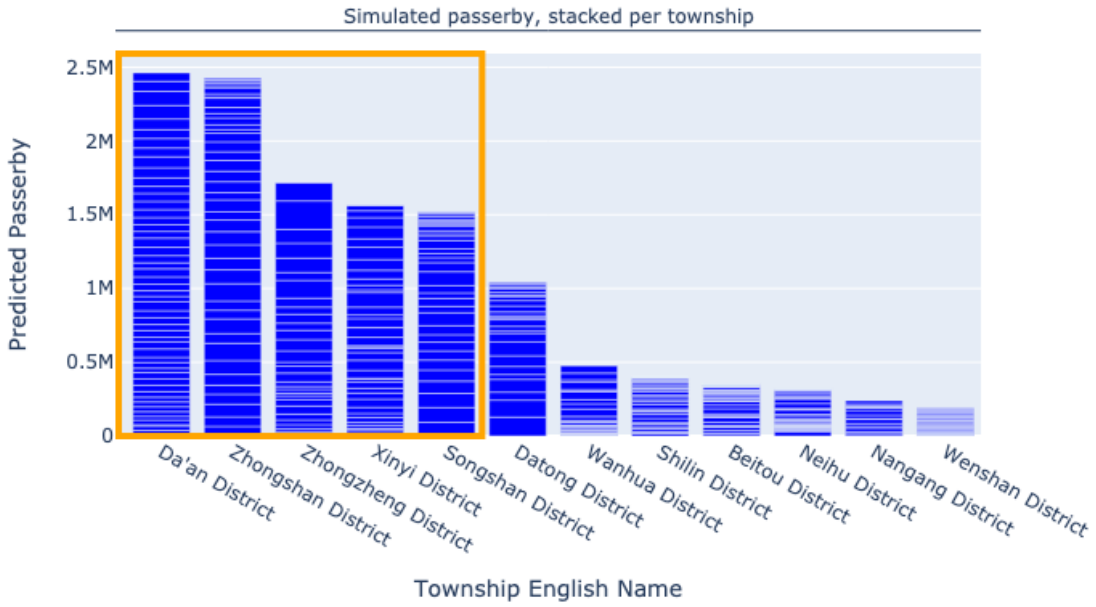
with result no more than 1 or less than 0

1.3.3 Data visualization

Here how is the data looks like.

Most of passerby are in Taipei mid-west area





1.4 Analysis

Analysis from the information are:

- **Most people activities is on Taipei mid-west Area.** They have about 3 times more passerby compare to other area.