Derive Community Structures from Taxi Flow Network

Multilevel Community Detection Algorithm

Basical Concept

Basically, the multilevel community detection algorithm aims to decompose a large network which has enormous vertex (may over millions) and edges into a set of highly inter-connected vertex, which can be also described as sub-units or communities. According to Blondel et al. (2008), traditional greedy algorithms have defects such as unsatisfactory modularity results, overfitting, low efficiency and so on. The multilevel community detection algorithm avoids these defects by introducing tricks which can balance the size of the communities being merged. In general, the tricks include segment, iteration, and aggregation. This algorithm can be also called BGLL or "Louvain method" (Blondel et al., 2023).

To measure the quality of partition results, modularity is always used as a quantitative measurement. It can scale the density of link inside communities as compared to links between communities (Blondel et al., 2008). Its definition in undirected weighted networks is

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j), \tag{1}$$

where A_{ij} represents the weight of the edge between i and j, $k_i = \sum_j A_{ij}$ is the sum of the weights of the edges attached to vertex i, c_i is the community to which vertex i is assigned, the $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise, and $m = \sum_{ij} A_{ij}$. What is more, the quality function has other four modifications that are suitable for different contribution models for inter-community edges or in the absence of edges (Blondel et al., 2023).

This algorithm can find high modularity partitions of large networks in a significant short time compared to other algorithms.

Algorithm

The step of community detection algorithms is intuitive and easy to implement (Blondel et al. ,2008). It can be divided into two phases. The first phase is to calculate the gain of modularity of the adjacent nodes of each node one by one and integrate them into one community if the gain is both positive and maximum. And then recursively execute this step for all nodes until no further improvement can be achieved, which means no individual move can improve the modularity. Phase one can also be modified by removing vertices that have little or no impact on the results or starting with communities which do not only contain a single vertex (Blondel et al., 2023). The formulation is,

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right], \quad (2)$$

where $\sum_{in}$ is the sum of the weights of the links inside C, $\sum_{tot}$ is the sum of the weights of the links incident to nodes in a community C, k_i is the sum of the weights of the links incident to node i, $k_{i,in}$ is the sum of the weights of the links from i to nodes in C and m is the sum of the weights of all the links in the network.

The second phase is to integrate the communities formulated in phase one as new nodes, and the weights of the links between them are the sum of the weights of the links between nodes in the corresponding two communities. And the sum of the weights of the link between the node of same community in phase one are calculated as self-loops. And then, calculate the new graph using the method in phase one again to formulate a new community. Doing these two phases iteratively until the modularity reaches the highest can get the final community.

According to Blondel et al.'s (2023) new research, during iterations, vertices that have not moved for several iterations or those that belong to stable communities can be ignored. Also, if a vertex has recently moved to a new community, its neighbors not in that community are considered likely candidates for movement. And central vertices (seed vertices) are prioritized for moving non-seed vertices to neighboring communities that maximize modularity gain. In some implementations, vertices are considered in a random order, allowing for exploration of diverse solutions. And vertices can be processed in order of decreasing centrality, which can enhance results and speed up the algorithm.

The steps of the algorithm are straightforward and have several advantages. It can operate without the need for supervision; It exhibits linear complexity on typical and sparse data; It simplifies the calculation of potential modularity gains; It circumvents the resolution limit problem associated with modularity optimization, allowing for the identification of smaller communities; It provides a decomposition of the network into communities at various levels, enabling users to explore the network's structure with different resolutions; And the intermediate results can be significant, offering insights into the network's organization before final aggregation.

The Results of The Communities

There are some special steps in data processing. Because some trips ended on the second day, the time of every trip will be counted based on the pick-up time. And some data’s intersection cannot pair with taxi zones, so these data were deleted directly.

The Whole Year Trips Occurred During 07:00 – 09:00

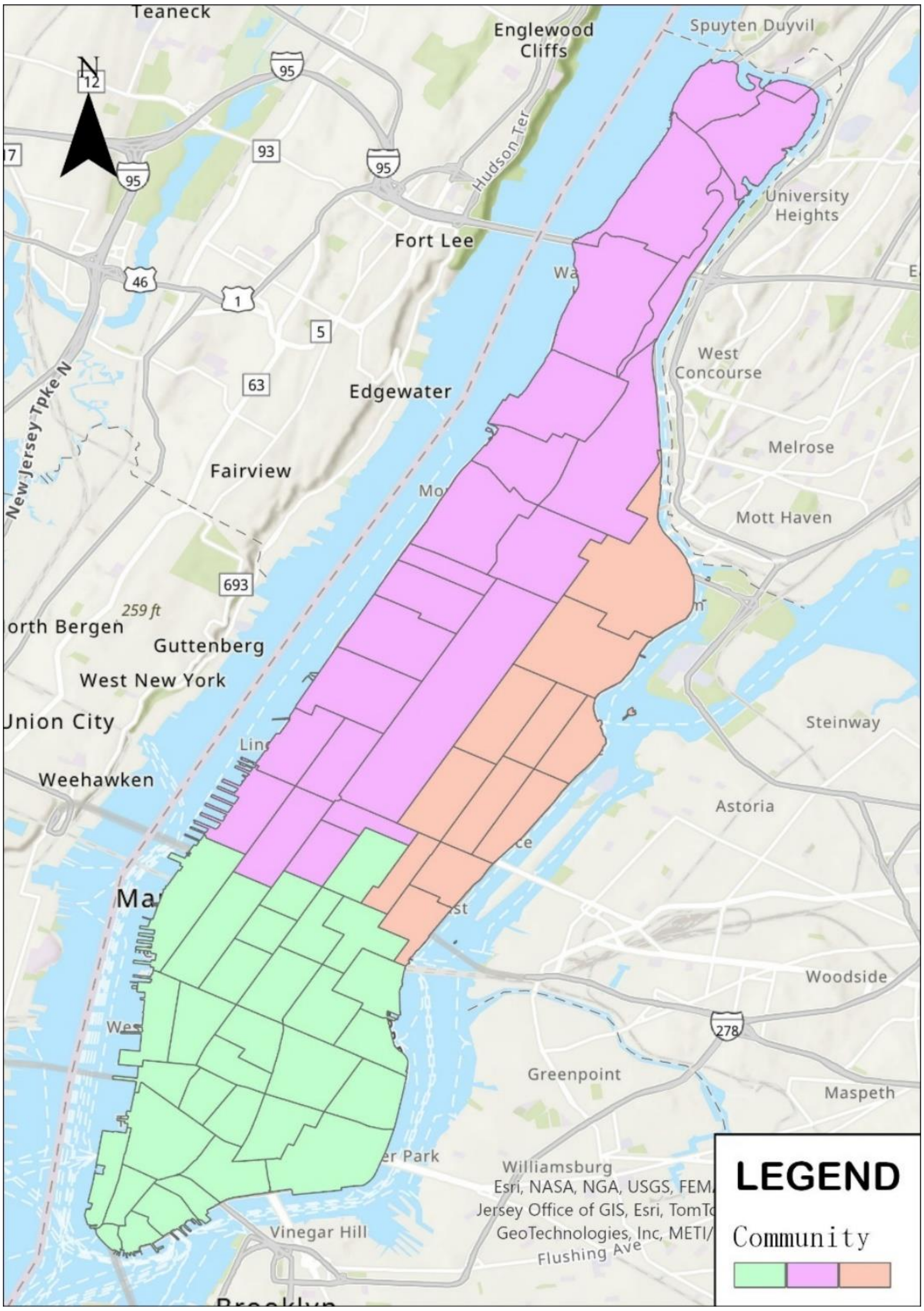

Fig 1. The Community Structures of Whole Year Trips Occurred During 07:00 – 09:00

The Whole Year Trips Occurred During 16:00 – 18:00

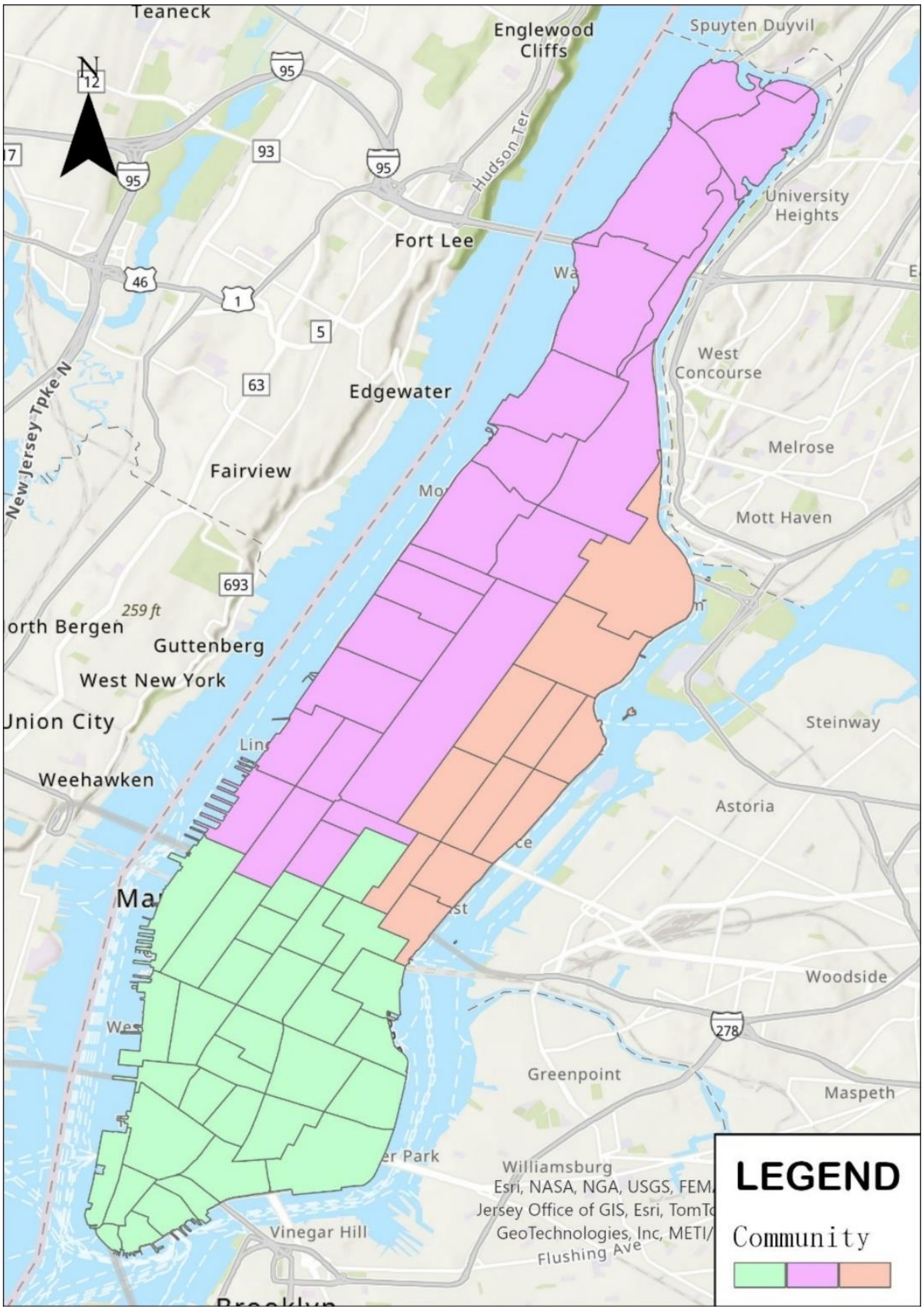

Fig 2. The Community Structures of Whole Year Trips Occurred During 16:00 – 18:00

From the result, it can be found that the community pattern of the whole year trips occurred during the 07:00 to 09:00 is same as the pattern of 16:00 to 18:00. This mainly because that these two time periods are working hours and closing hours respectively, and most people would go back wherever they come from.

The Trips Occurred in January

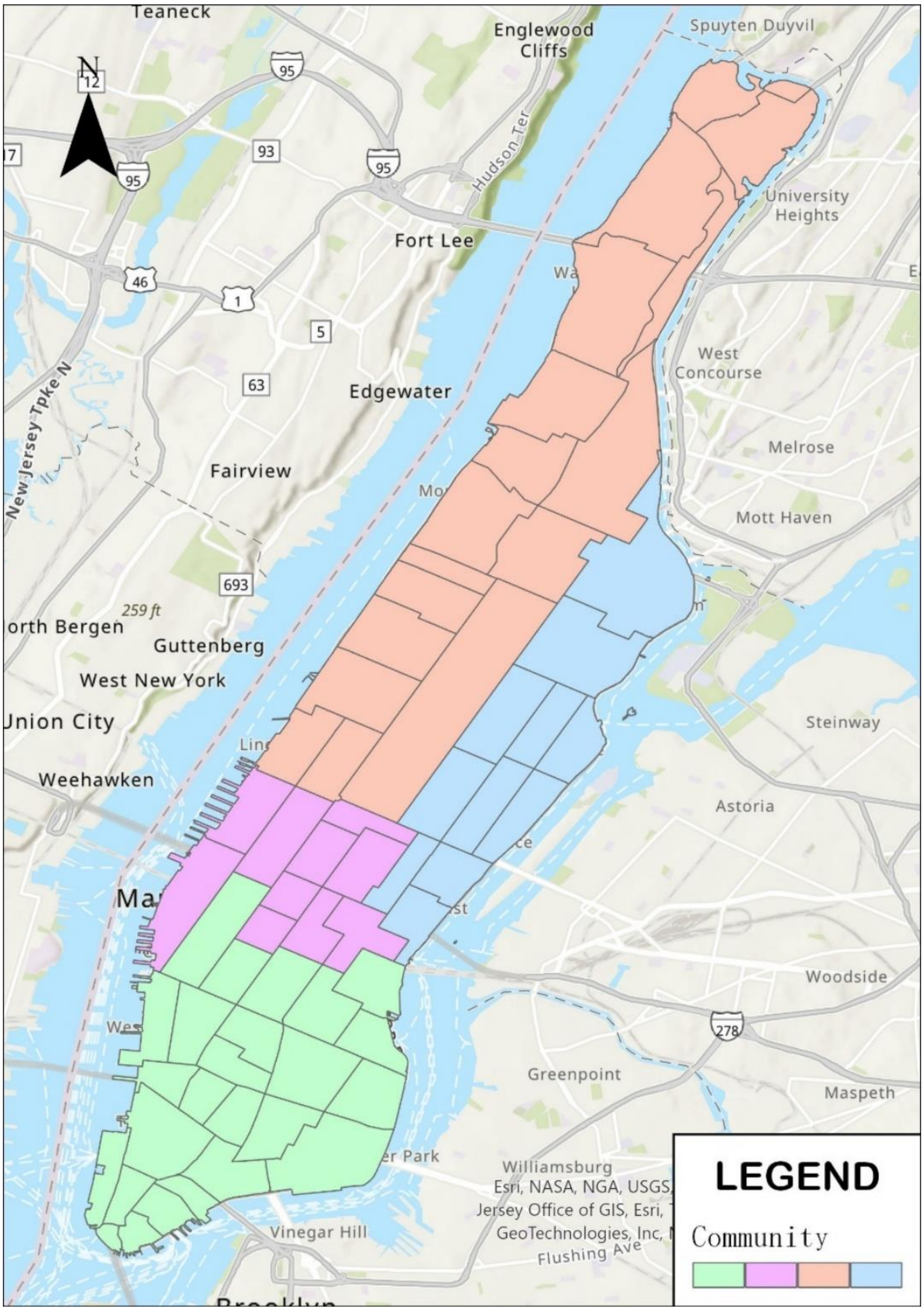

Fig 3. The Community Structures of Trips Occurred in January

The Trips Occurred in February

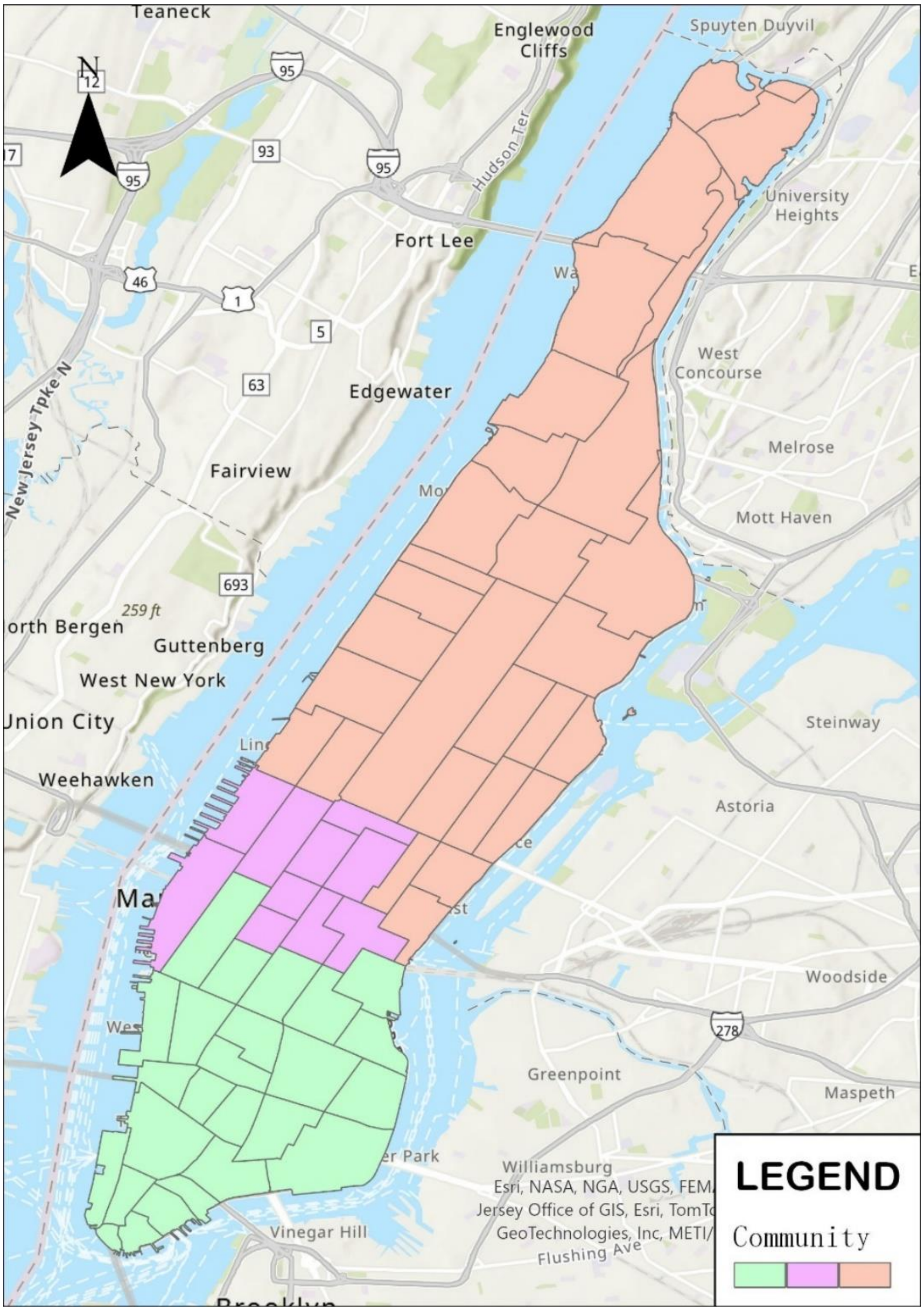

Fig 4. The Community Structures of Trips Occurred in February

The Trips Occurred in March

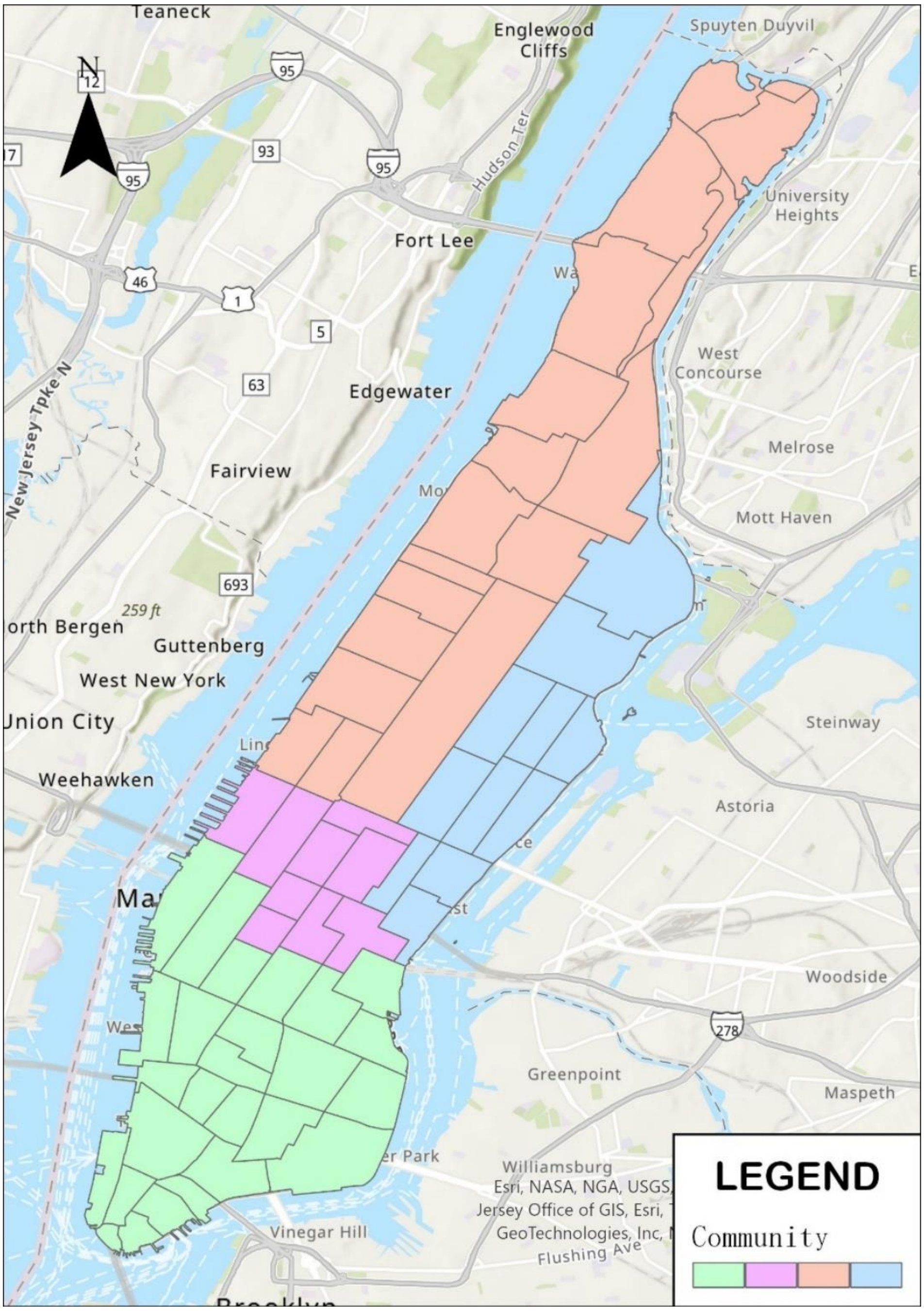

Fig 5. The Community Structures of Trips Occurred in March

The Trips Occurred in April

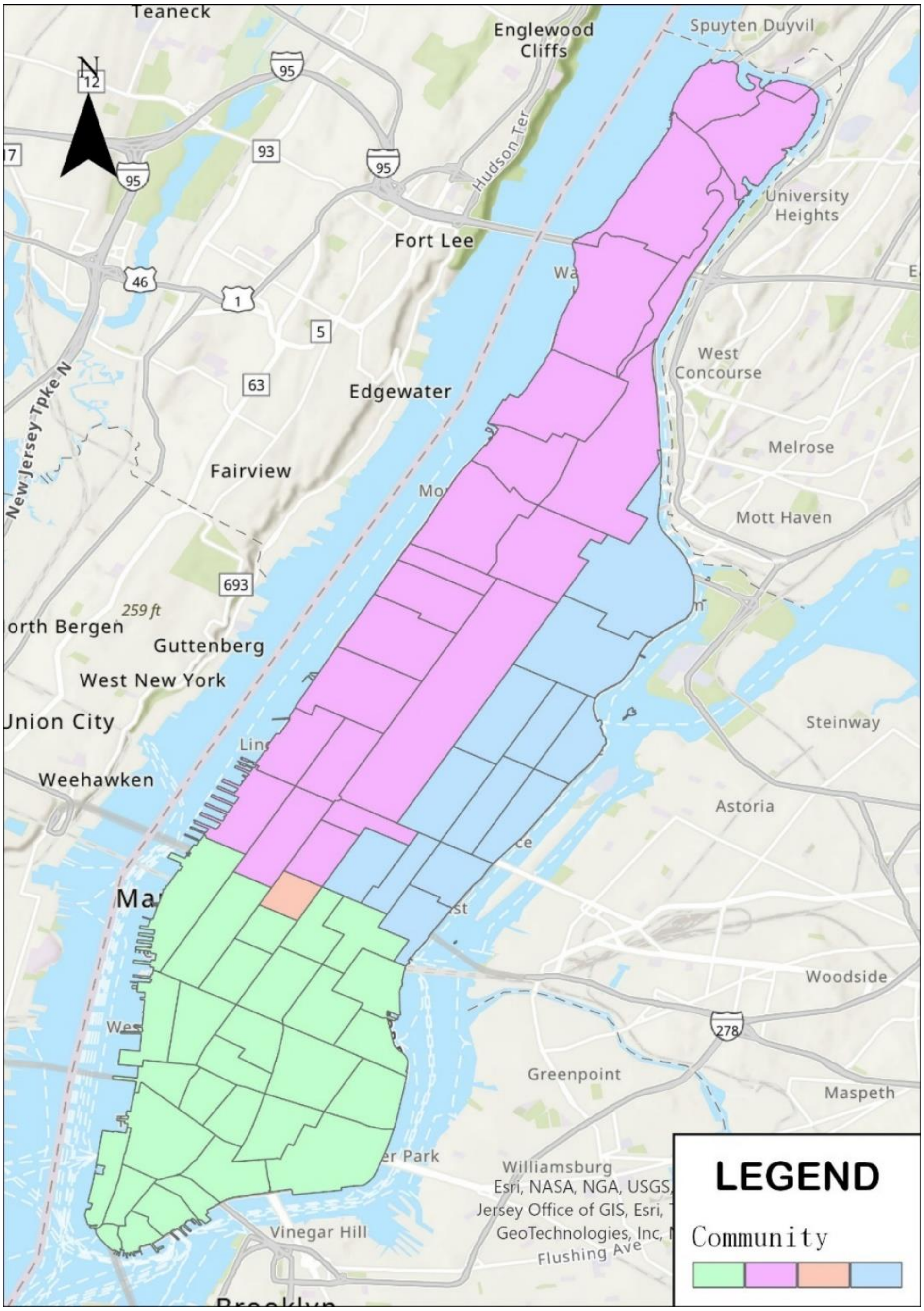

Fig 6. The Community Structures of Trips Occurred in April

The Trips Occurred in May

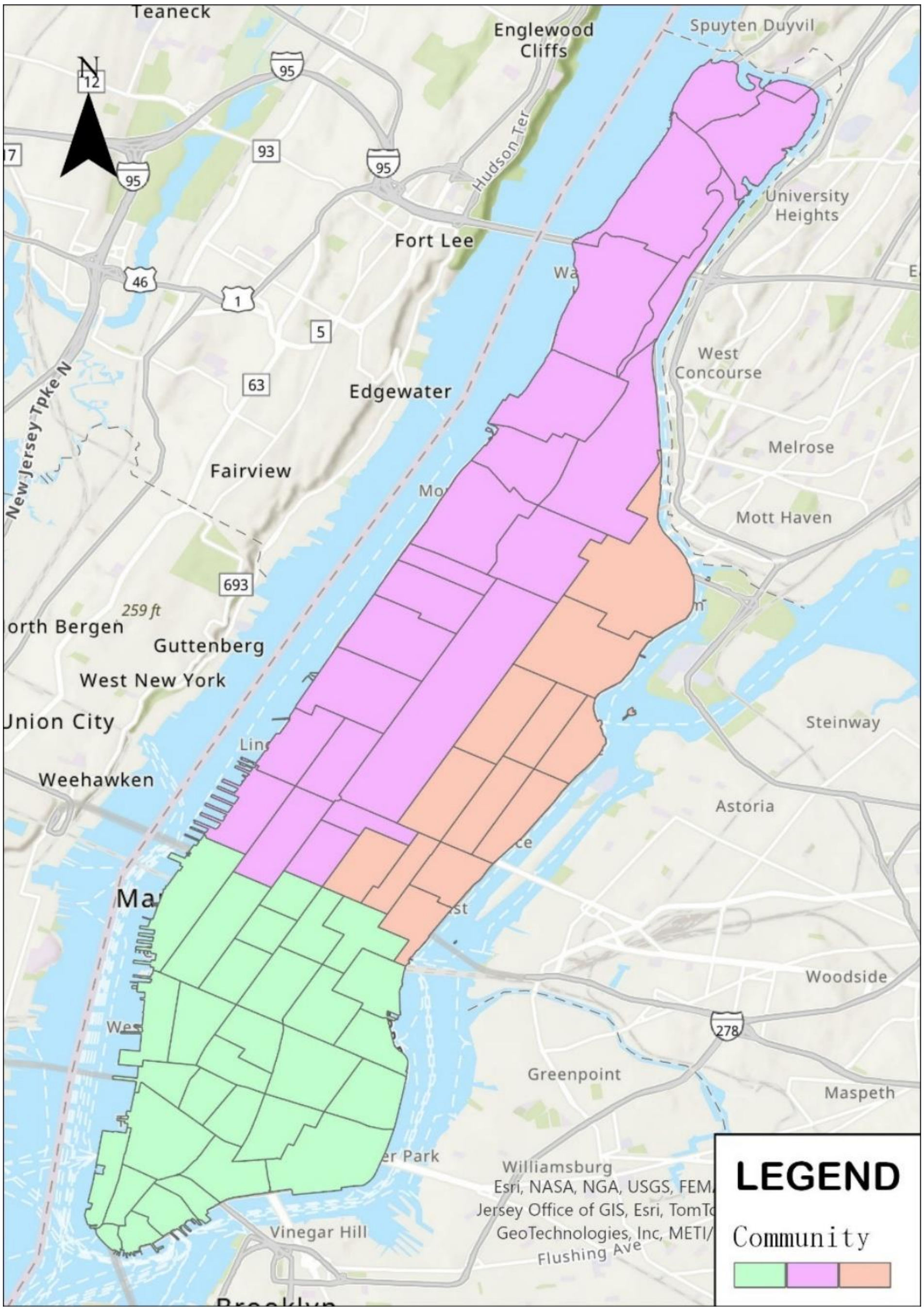

Fig 7. The Community Structures of Trips Occurred in May

The Trips Occurred in June

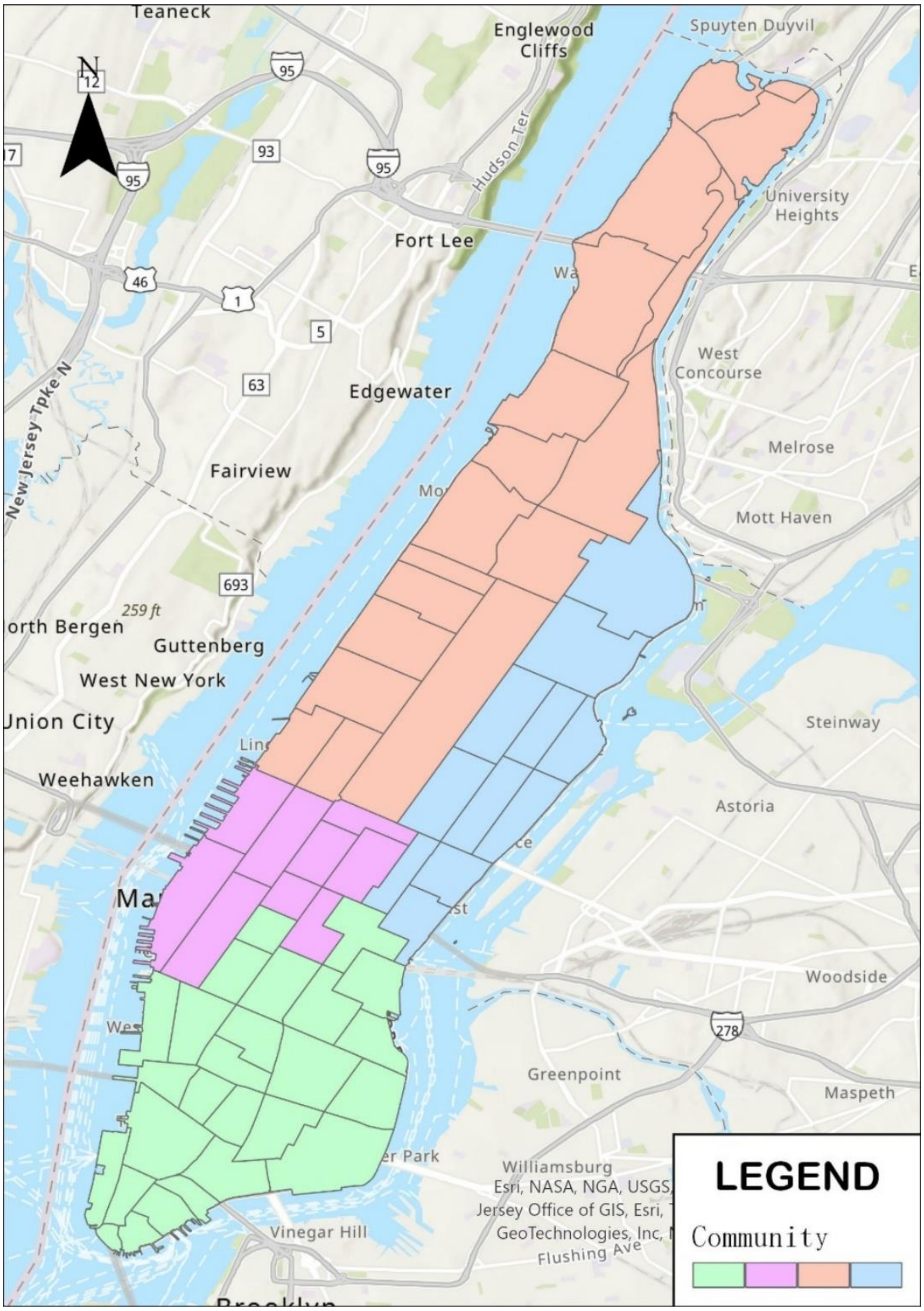

Fig 8. The Community Structures of Trips Occurred in June

The Trips Occurred in July

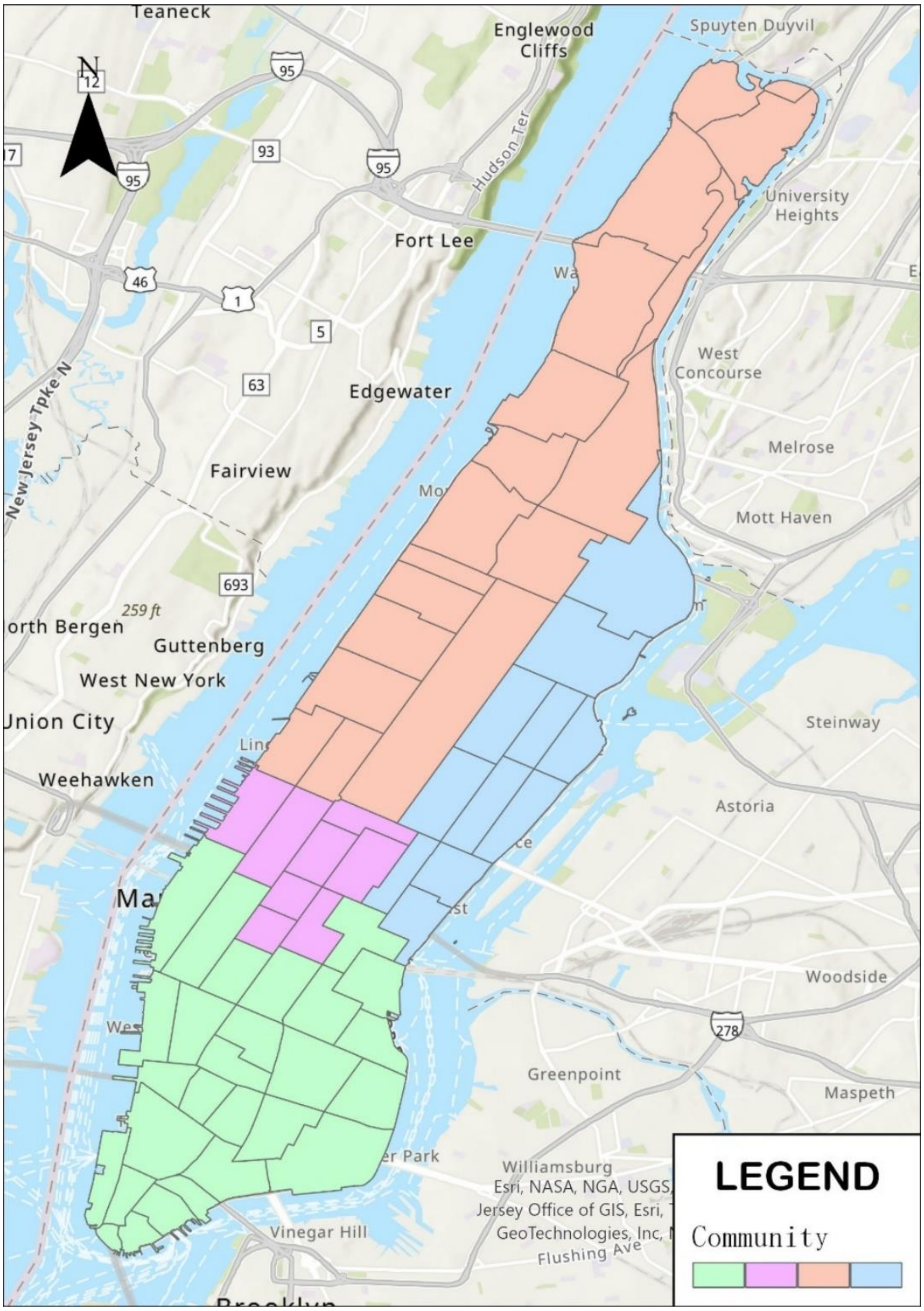

Fig 9. The Community Structures of Trips Occurred in July

The Trips Occurred in August

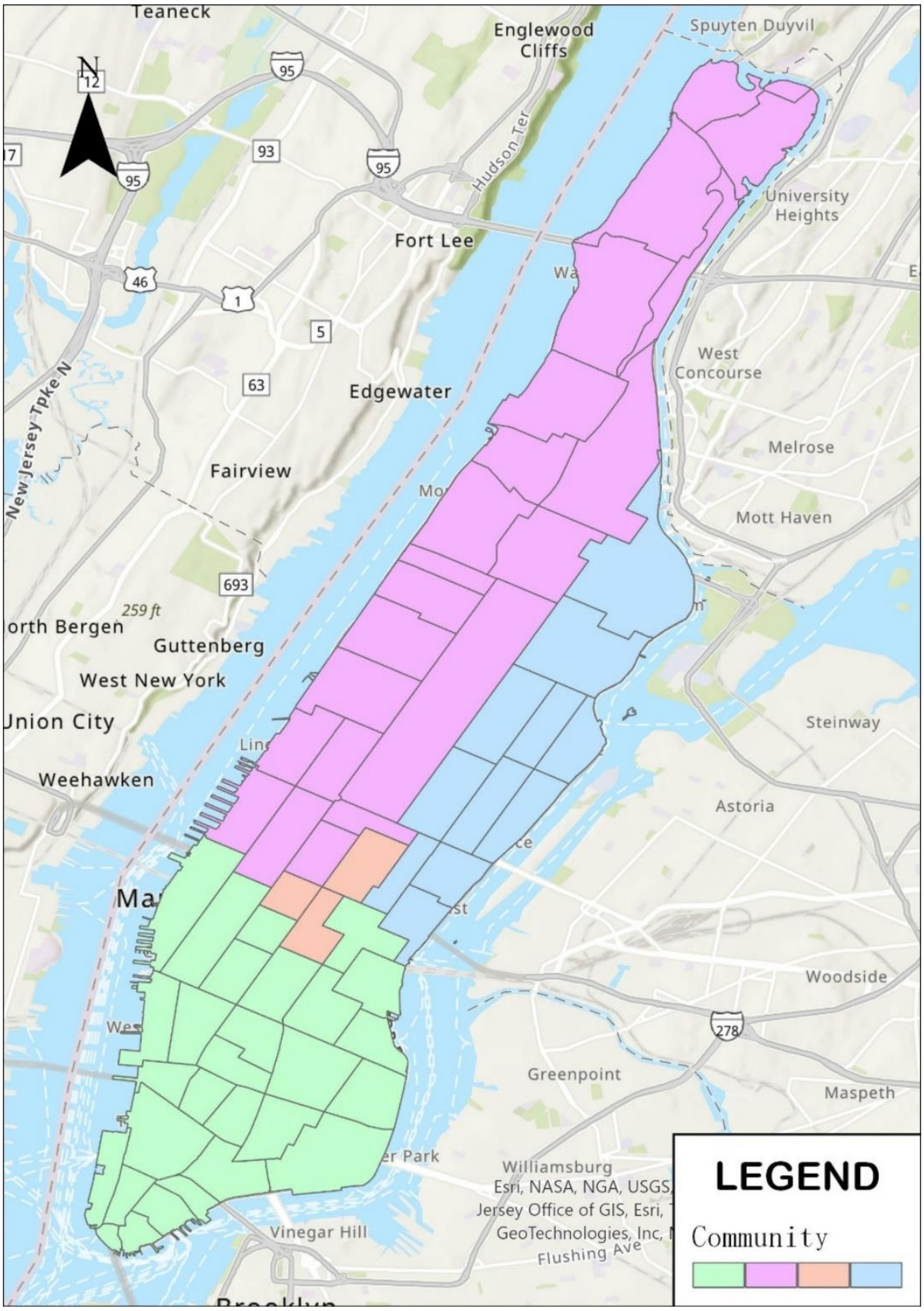

Fig 10. The Community Structures of Trips Occurred in August

The Trips Occurred in September

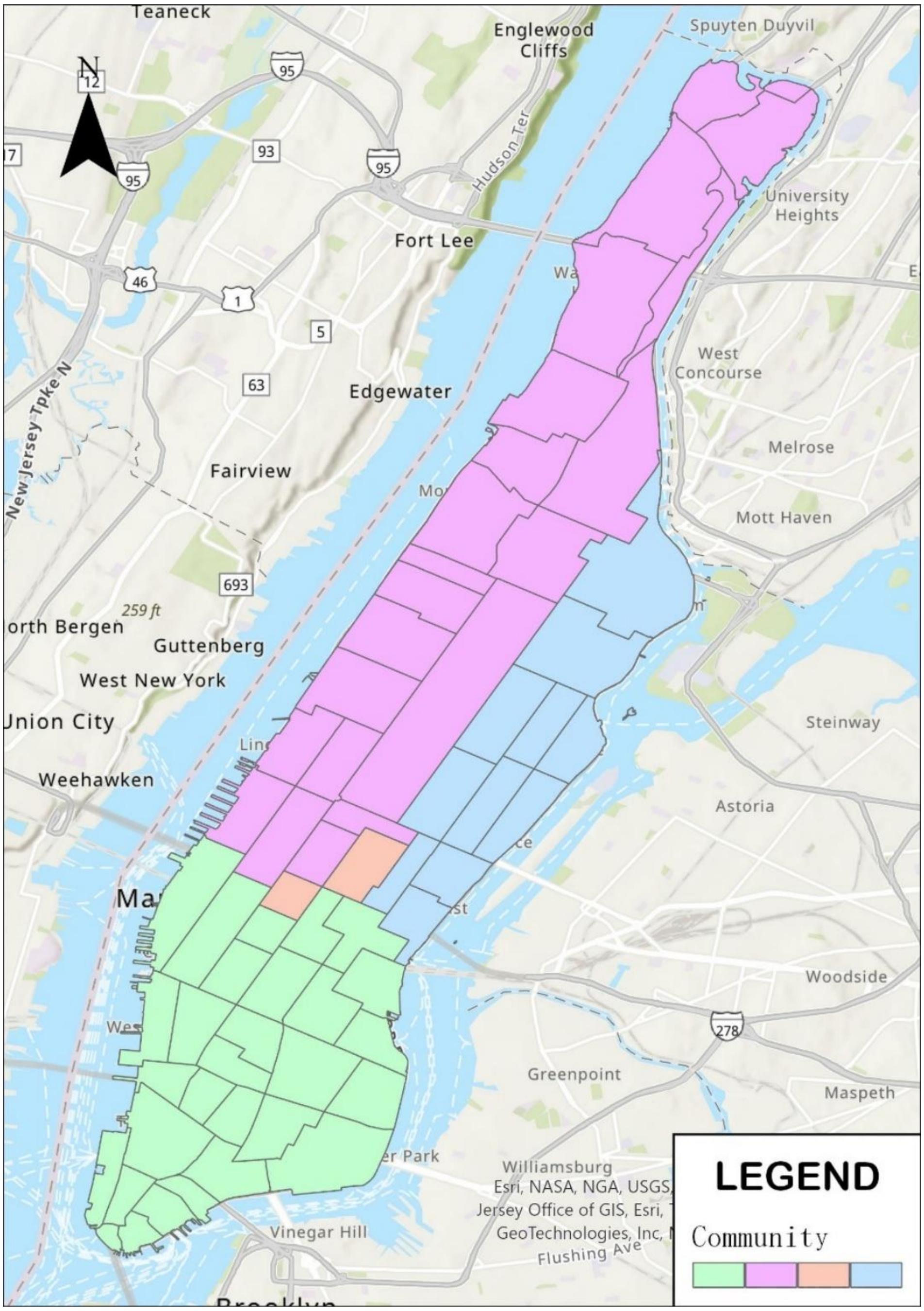

Fig 11. The Community Structures of Trips Occurred in September

The Trips Occurred in October

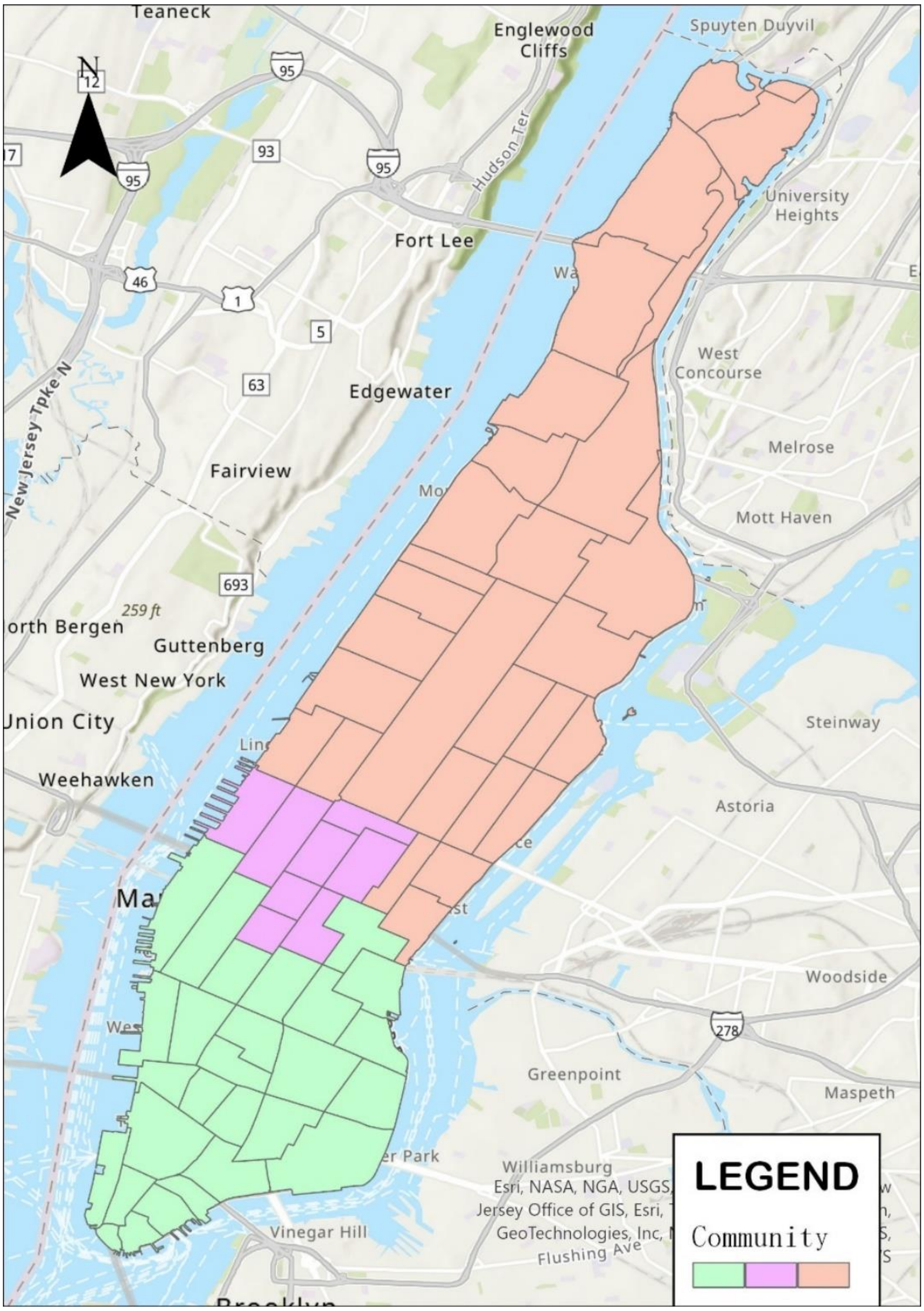

Fig 12. The Community Structures of Trips Occurred in October

The Trips Occurred in November

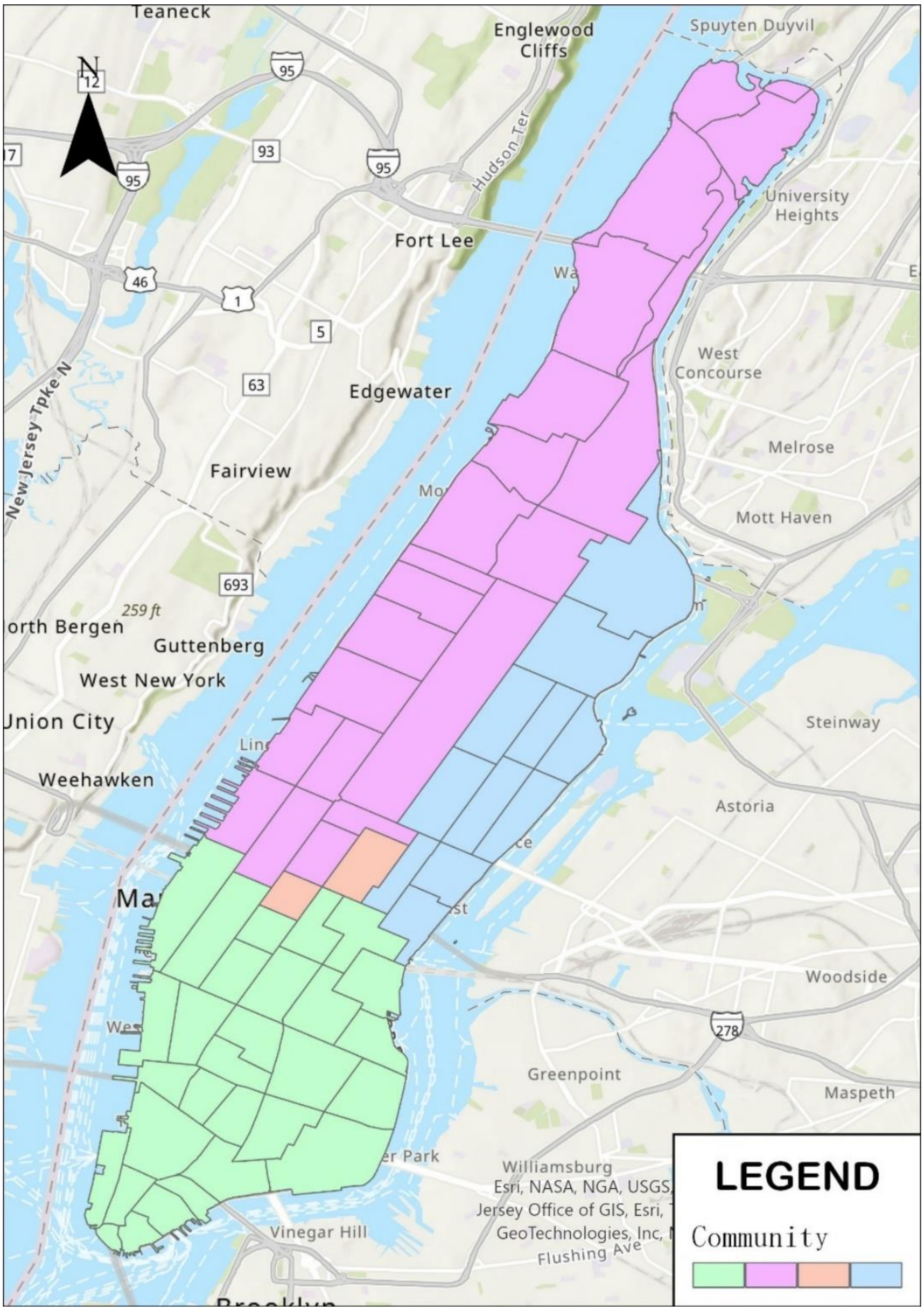

Fig 13. The Community Structures of Trips Occurred in November

The Trips Occurred in December

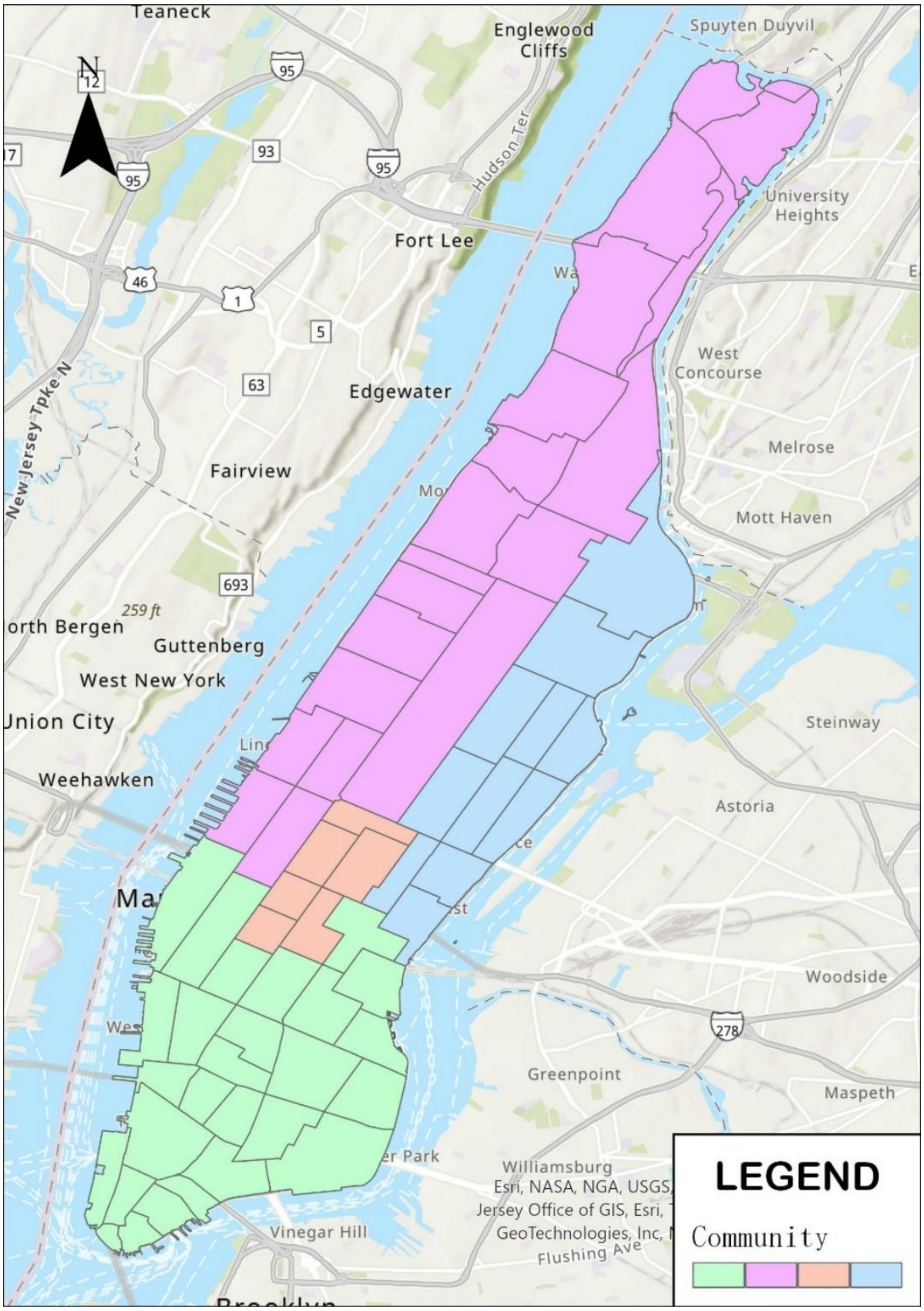

Fig 14. The Community Structures of Trips Occurred in December

From the trips that occurred in January to December, it can be found that the community patterns only have slight differences. The north part where is residential area is divided into a same community area as time goes on, and the south part where is financial district is also always divided into another same community as time goes on. This may be because people who live or work in these areas usually do not change their commuting patterns frequently. The south north area where is the transition region between residential areas and financial districts is usually divided into a same community as time goes on, but merge into the north area in February and October, this may be because some slight change in tourist flow.

However, the center area where has lots of scenic spots has different divided patterns as time goes on. This may because the taxi routes are affected by the tourist, the number of tourists will vary in different month due to the public vacations, and the destinations of tourists will vary in different month due to the weather conditions and time sensitive attractions such as river.

Reference

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10), P10008. https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008

Blondel, V., Guillaume, J. L., & Lambiotte, R. (2023). Fast unfolding of communities in large networks: 15 years later. *arXiv preprint arXiv:2311.06047*. https://doi.org/10.48550/arXiv.2311.06047