Aggregating taxi rides for microtransit

Code can be found here: https://github.com/sz5120/VIA-Data-Challenge/blob/main/ride_aggregation.ipynb

Q1. An algorithm/metric to assess the potential efficiency of aggregating rides, stating realistic assumptions and simplifications.

As a first-order metric for assessing the potential efficiency of ride aggregation, use the fraction of aggregated rides relative to the total number of rides, which I will refer to as the aggregation rate.

Consider three forms of ride aggregation:
- a large single bus that picks up and drops off many passengers along the way, running on a regular schedule depending on time, location, and day
- smaller vans that pick up and drop off passengers along the route on demand
- smaller vans that pick up all passengers in the same place, and drop them off in the same place.

The third method will be used here.

In order to implement the aggregation potential for this metric, we make several assumptions and simplifications. We assume that all rides within 500m of each other can be picked up at some average location and that passengers will walk to this location; all rides will be dropped off within 800m at some average location and all passengers will walk from this drop-off location to their destination; all rides are picked up within a 5-minute window, and all vans have the same maximum capacity. We will not consider fare or fuel efficiency, passenger utility, cost of operating the vans, or profitability of these aggregated rides.

Possible vehicle aggregations are determined in two main steps. First, individual rides will be clustered based on pick-up location, pick-up time, and drop-off location (HDBSCAN). Second, individual rides within each cluster will be assigned to a van based on the above constraints. Each van will be considered a valid aggregated ride.

In order to calculate some metrics, rides with characteristics which lie outside the 90th percentile (>2 million rides remain) are excluded. A constrained total sample size of rides (80,000 rides per day) with each day divided into 30-minute buckets, with each bucket holding maximum of 3000 samples for the clustering analysis. A maximum cluster size of 20 vehicles is assigned for the clustering analysis, and allow any van that can aggregate at least 2 individual vehicles.

It is also reasonable to expect that individual rides are more successful to be aggregated when and where there is a higher density of individual rides. In addition to spatial variation, it is expected there will be temporal variation based on the broad schedule of human activity (see Fig. 1a, Fig 2). Manhattan neighbourhoods are strongly defined by the activities that occur within them. For more meaningful spatial analysis, rides will by labelled census tracts and the neighbourhood they are categorised in. As seen in Fig 2, there are clear patterns in pick-up and drop-off times based on neighbourhood and day of the week. For example, it is likely that weekday rides dropped off in Midtown neighbourhoods between 7 and 9 am are the morning rush-hour from the start of the work-day, while those picked up more broadly 5-7pm correspond to the end of the work-day. Drop-offs in Midtown and Lincoln Square around 2 PM likely correspond to Saturday matinee shows.

Q2. Implement method and evaluate it using Manhattan's taxi data from the first full week of June 2013. Discuss how it would scale with more data.

Before evaluating ride aggregation potential, a relevant sample from the 2013 taxi data covering all five New York City boroughs is selected. The first full Monday-to-Sunday week in June 2013 is from June 3 to June 9. (rides picked up between 2013-06-03 00:00 and 2013-06-09 23:59). For the purpose of this study, consider only rides that initiate and finish in Manhattan with a rate code of 1 (regular street fare). Only the base fare information is used, excluding fees, tips, and tolls. Outliers (90^{th} percentile) of rides based on trip length, distance, and speed are excluded. The filtered data set evaluated contains 2.2 million individual rides.

I choose a HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) clustering algorithm as it can handle outliers, in comparison to e.g. k-means centroid-based clustering. Compared to DBSCAN, it is more effective for data with inhomogeneous density (such as taxi rides in Manhattan). It also does not require a pre-defined number of clusters. The default HDBSCAN method from scikit-learn is used on clustering features of pickup time, pickup latitude and longitude, and drop-off latitude and longitude. Using a 30-minute range of pickup time allows for approximately equal weight with distance in tolerance levels when scaled. We set parameters of minimum cluster size of 2 (including the initial point) and maximum cluster size of 20.

Within each cluster, then attempt to aggregate individual rides into pooled vans. For every valid van, all individual rides must have been picked up within 5 minutes, picked up within 500m, and dropped off within 800m (assume it is possible for multiple drop-off locations). Assume the trip distance and trip time of the van to be the average of those of all individual rides. While this is neither the most efficient method for ride assignment nor the most optimal combination, it will function as a minimal working model, and can also be tuned for larger minimum and maximum sample size and vehicle capacity (e.g. for bus routes).

With larger datasets, the primary increase in complexity would be in the clustering analysis. As the amount of data increases, the number of clusters will presumably also increase if the maximum cluster size is kept the same. While the per-cluster ride aggregation would scale poorly with size, with our constraints on features and small epsilon, HDBSCAN (and DBSCAN) will have a reasonable run time (see various sk-learn documentation) for most reasonable datasets.

However, using a spatial indexing system such as H3 would improve performance. (This method has been applied to NYC taxi data, but I learned about H3 too late to implement it here.)

The code can be found here: *https://github.com/sz5120/VIA-Data-Challenge/blob/main/ride_aggregation.ipynb*

Q3. Visualizations of how efficiency varies with time and location. Discuss any potential business implications.

See attached figures for analysis of efficiency as a function of time and location.

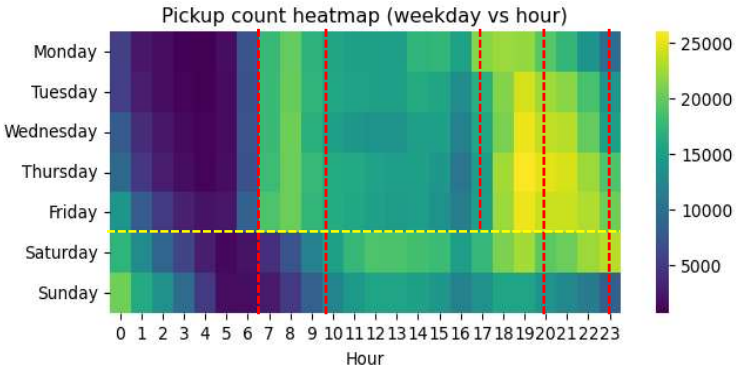
In brief summary: the aggregation rate is higher when there are more individual rides, up to ~18%. Rides that can be aggregated are often dropped off in Midtown in the mornings and picked-up in the evenings and strongly follows work day patterns, as well as cultural activity patterns. Within Manhattan, shorter routes within Lower Manhattan and Midtown are more likely candidates for aggregation. Most rides (75th percentile) are aggregated into vans with 6 or fewer passengers, or 4 or fewer passengers depending on neighbourhood.

Business implications:

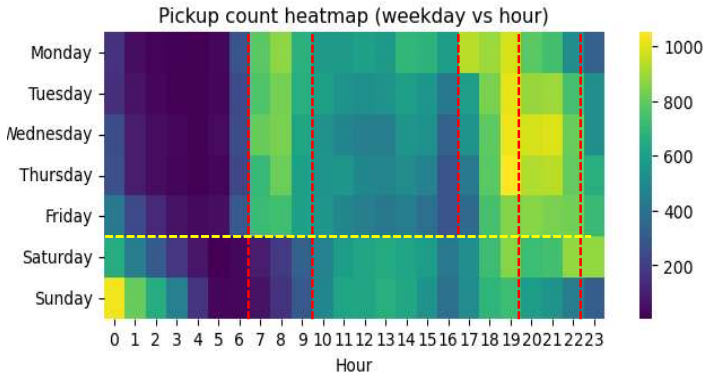
- Routes where ride aggregation can be successful
- Vehicle capacity for aggregating rides
- Determine potential profitability given operating costs
- The same algorithm and metric can be tested with larger minimum cluster sizes and pickup windows to evaluate the efficiency of e.g. bus or shuttle routes
- Time and location to deploy pooling vehicles, for if a rideshare is requested
- Possible routes include:
    - Those departing from Midtown on Monday to Thursday between 4-7PM
    - Upper East Side to Midtown on Monday to Friday as well as lower Upper West Side
    - Certain neighbourhoods also have high aggregation efficiency Friday-Sunday from 11PM-2AM.

Some further notes:

Given these are individual rides for taxis, and that Manhattan is well served by fixed bus routes and an extensive subway system, the large portion of rides that can be reasonably aggregated along a specific route is not contained within this dataset. I also consider that the demographic of those who will take taxis are more likely to be connected to the Financial District than, e.g. a university district.



Individual rides pick-up times segmented by weekday and hour. Dashed lines mark out regions of ridership patterns.



Aggregated vans pick-up times, segmented by weekday and hour with the same guide lines.

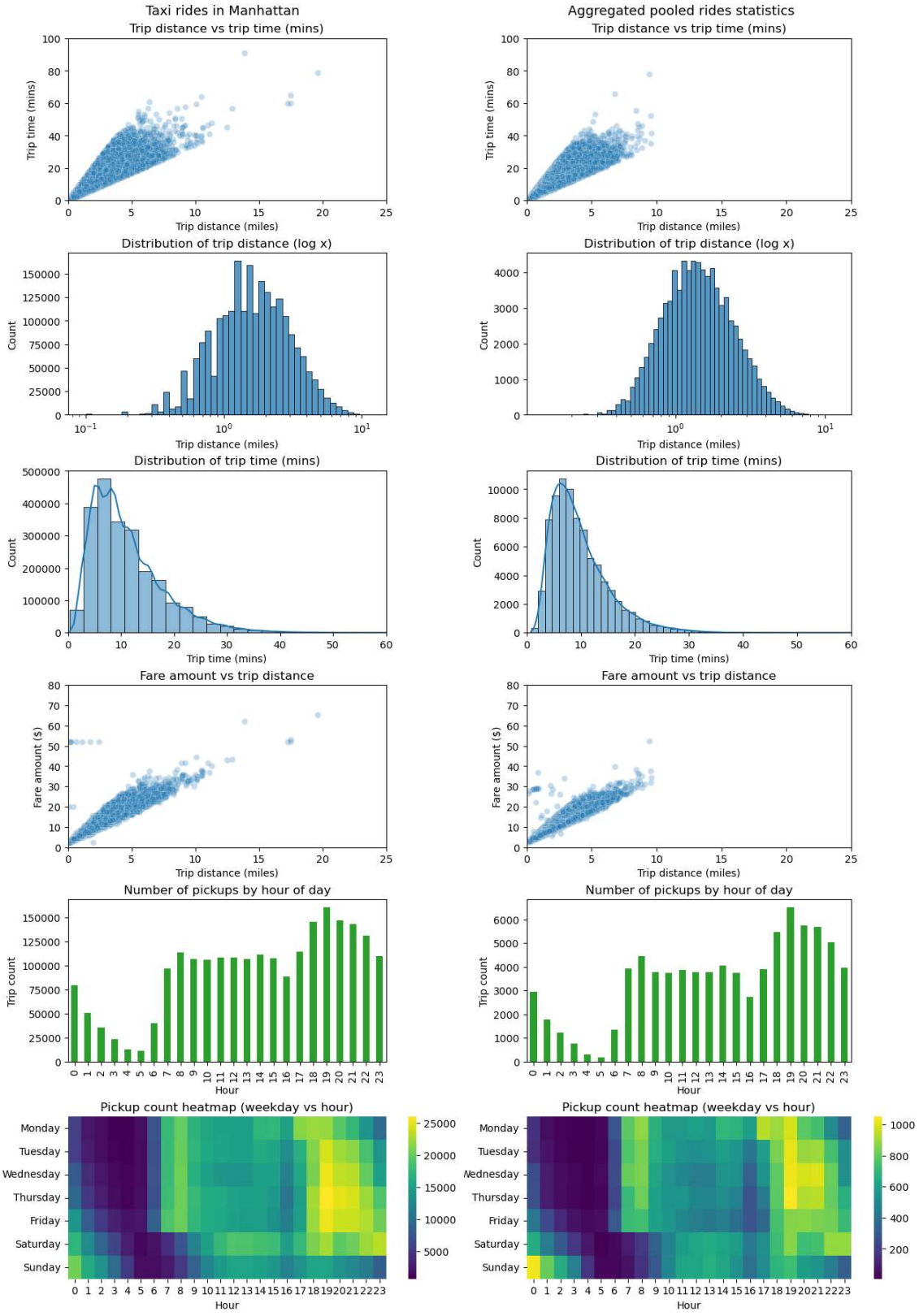


Figure 1. Overview of characteristics of taxi rides in the first week in June 2013 in Manhattan. a) Individual taxi rides. b) Aggregated van rides.

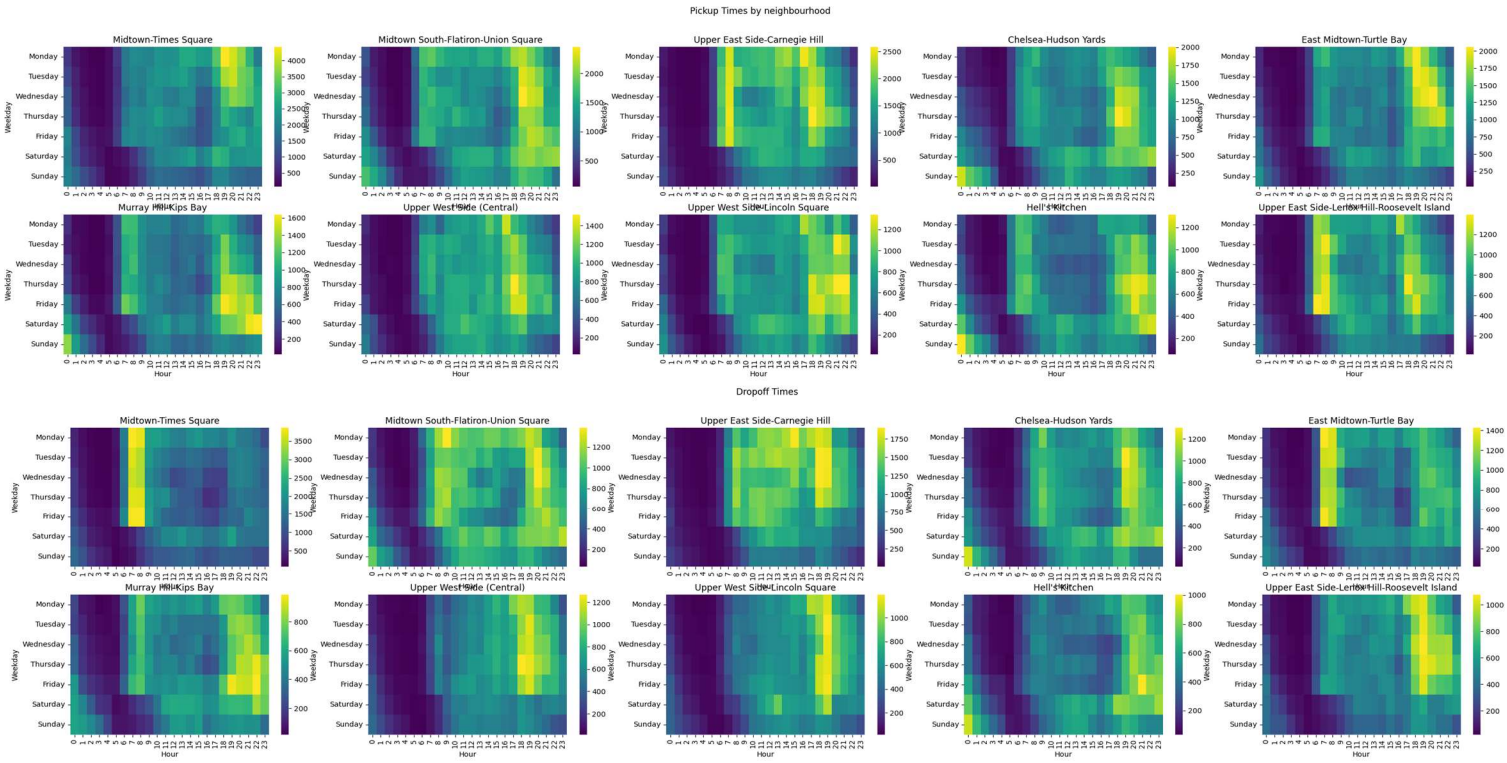


Figure 2. Pickup and Drop-off times for 10 Neighbourhoods with the most individual pick-ups. Midtown neighbourhoods strongly follow the work week and day schedule. Neighbourhoods with morning pickups and evening drop-offs show a more residential pattern.

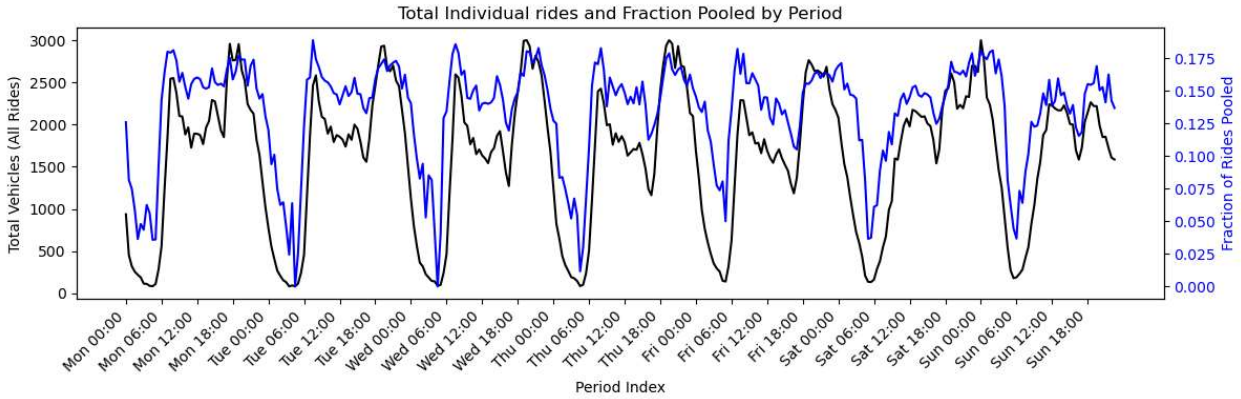


Figure 3. The fraction of rides successfully aggregated strongly follows the total number of individual rides

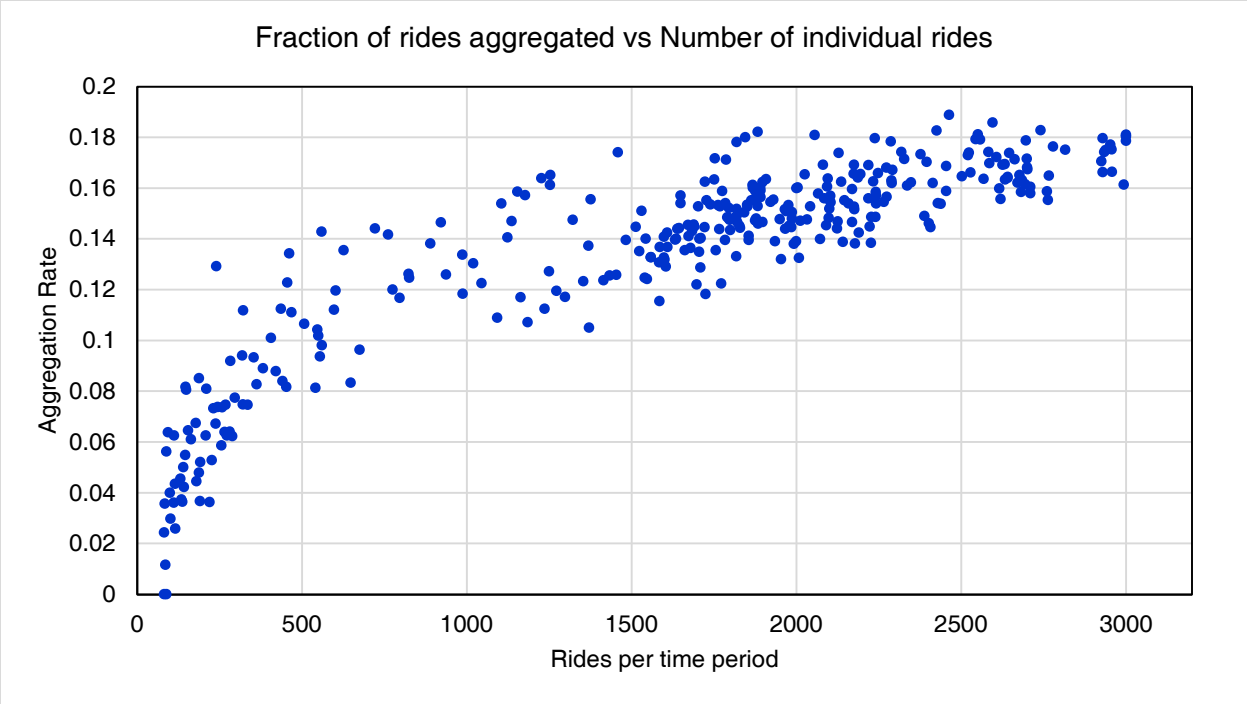


Figure 4. Fraction of rides that can be aggregated into a single van within a 30 minute period (aggregation rate), as the number of rides increases. The percentage of rides that can be aggregated converges to approximately 18%.

Feature	Van rides: median	Taxi rides: median	Van rides: mean	Taxi rides: mean	F-statistic	ANOVA p-value	t-statistic	t-test p-value
Trip time (s)	473.0000	720.000	524.487	810.311	104.319	1.291e-23	-13.173	1.963e-36
Trip distance(miles)	1.0300	1.740	1.173	2.059	112.818	2.465e-25	-15.853	8.923e-52
Fare ($)	7.0000	9.500	7.493	10.816	114.700	1.029e-25	-14.753	4.958e-45
Speed	7.9529	8.589	8.631	9.252	13.171	2.951e-04	-3.821	1.467e-04

Table 1. Comparison of trip characteristics of rides that were and were not able to be aggregated. Shorter rides were more likely to be successfully aggregated.

Individual Rides

Aggregated Vans

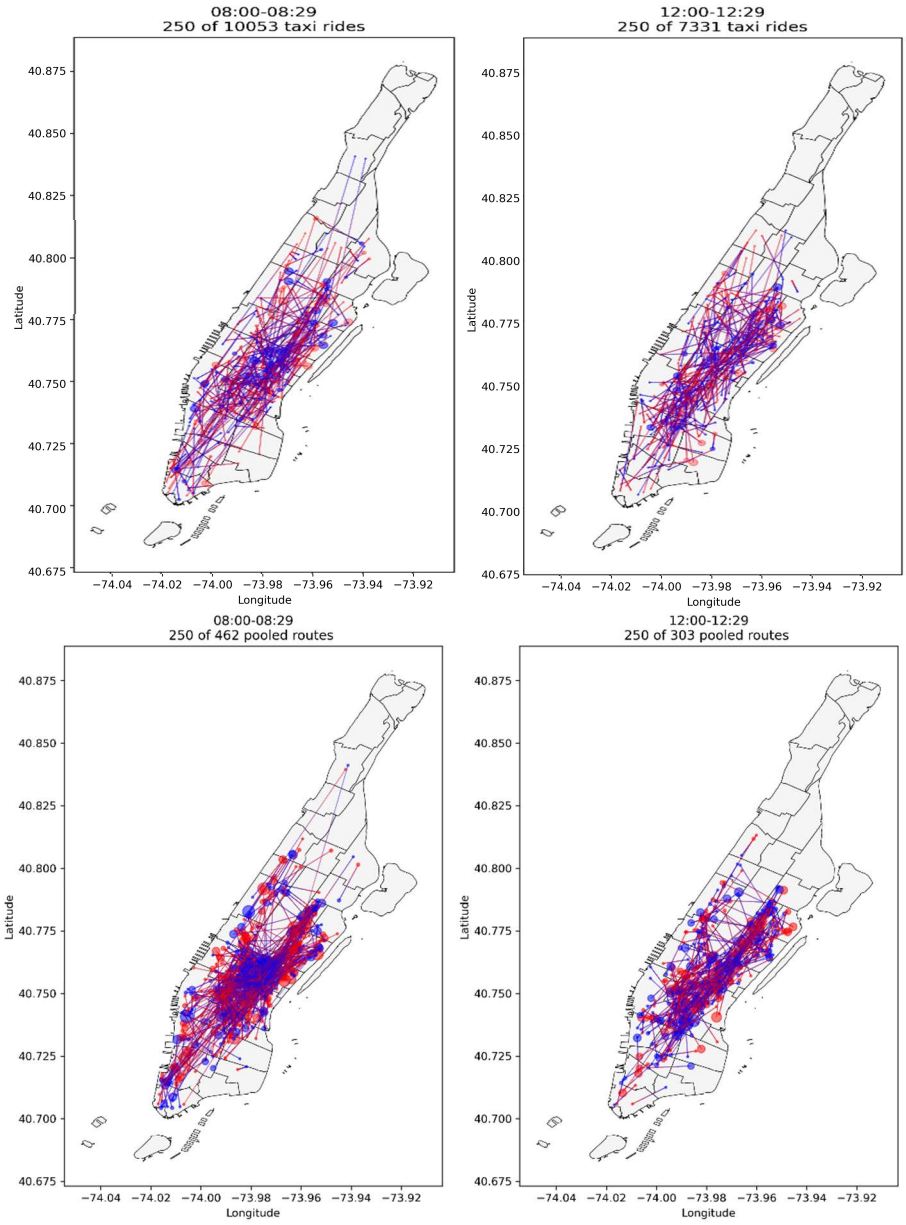


Figure 5. Comparison for a selection of 250 routes for individual (upper) and aggregated (lower) rides. Ride aggregation is more successful in regions with denser pickups and drop-offs, as well as for shorter rides. These two are likely correlated. Marker size indicates the number of passengers.

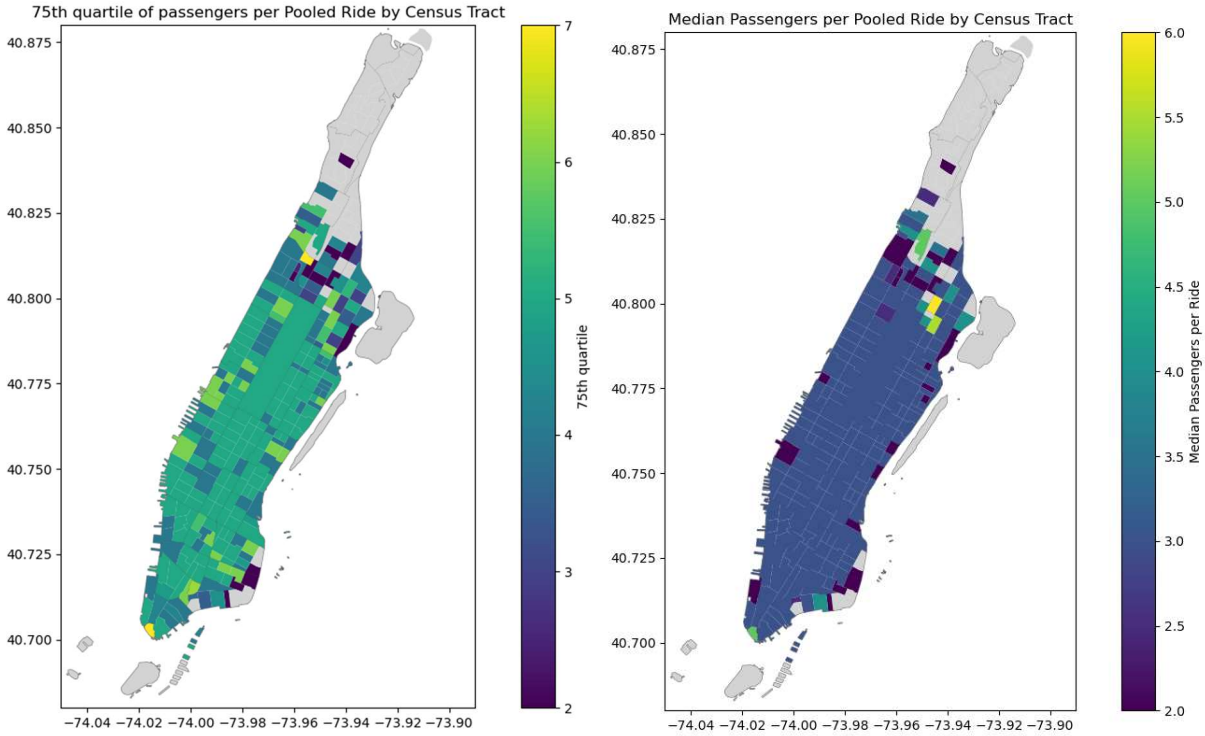
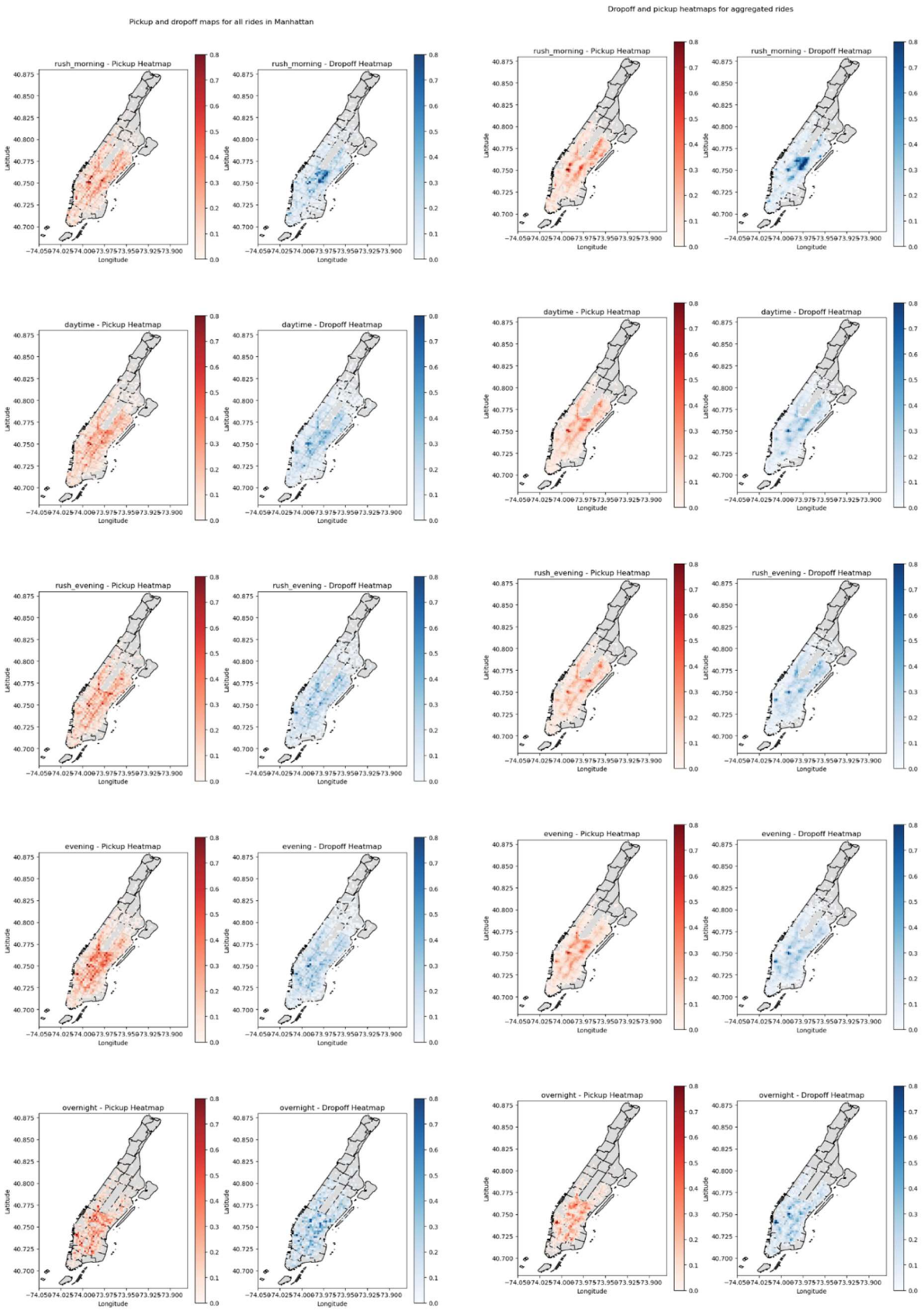
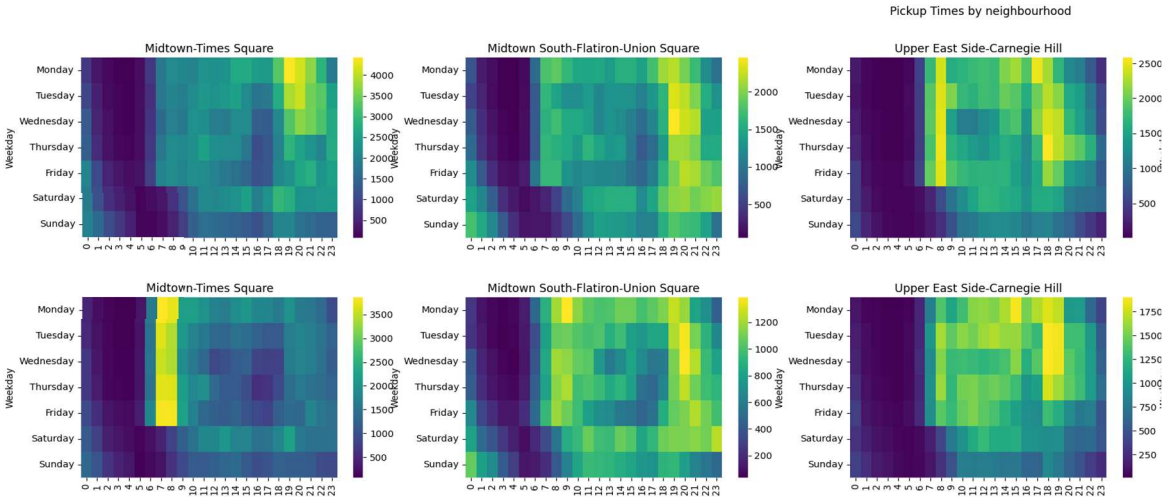


Figure 6. The median and 75th quartile of number of passengers per van in Manhattan. Most (75th percentile) of aggregated rides will have six or fewer passengers under these aggregation conditions. There is some variation by region.

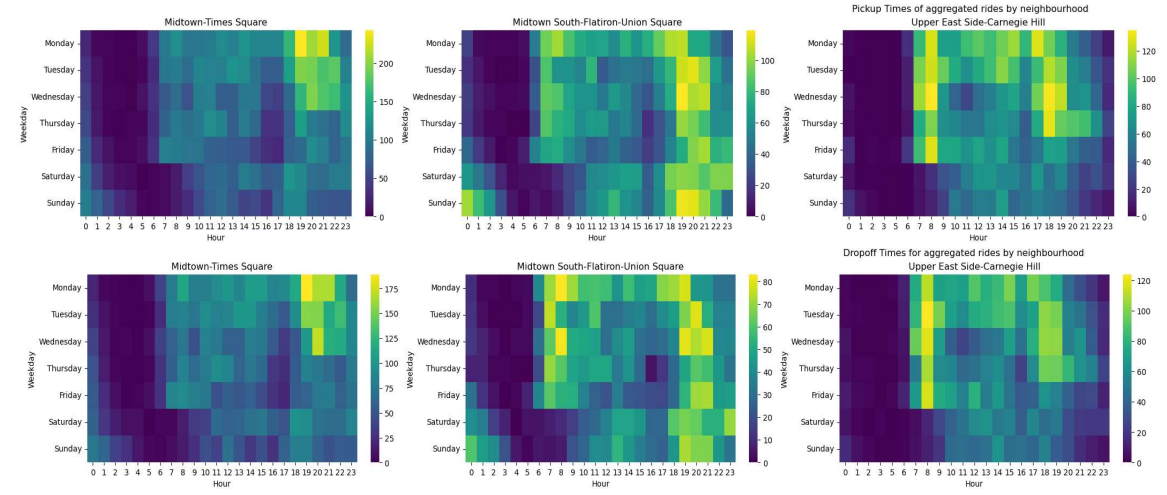


Spatial comparison for individual (left) and aggregated (right) rides for different time periods of pickup and drop-off locations. Aggregated rides tend to show fewer but denser hot spots, but follow similar spatial trends as total rides.

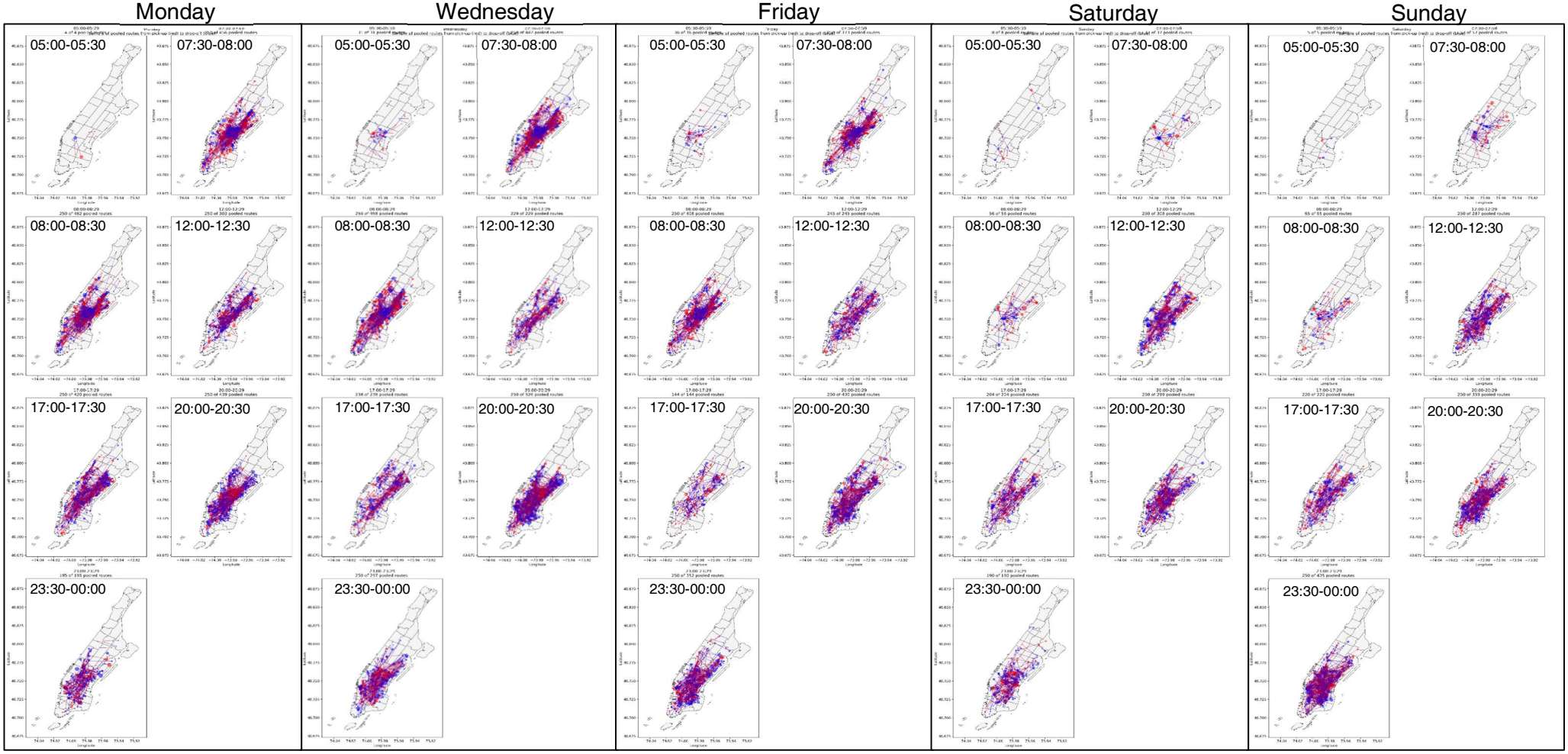
Individual Rides



Aggregated Vans



Comparison of pick-up and drop-offs for selection of neighbourhoods between individual rides and aggregated vans. The aggregated rides follow similar trends to individual rides.



Routes of aggregated vans with pickup locations marked as red and drop off as blue. Marker size denotes number of passengers. There is a strong dependence on location. Most aggregated routes tend to be in the lower half of Manhattan.

The absolute number and distribution of aggregated rides correspond strongly to the number of individual rides, as naively expected. The fraction of aggregated rides also follows this trend: the aggregation rate is higher when the number of total rides is higher (Figure 3).

Under the conditions where each van aggregates at minimum 2 rides that initiate within 5 minutes and 500m and end within 800m, the aggregation rate increases with absolute number of individual rides before tapering off to an aggregation rate of 18% (Figure 4.)

By comparing trip characteristics, rides that were successfully aggregated tended to be shorter in time and distance than rides in general (Table 1.) This is visualised in Figure 5 which shows a selection of all individual routes compared to aggregated routes. Ride aggregation is more successful in regions with denser pickups and drop-offs, as well as for shorter rides.