Infering the Night Life Hotspot in New York City from Taxi Trip Data

Wenhao Wu (UCD ID: 998587583)

***Abstract*—In this work, we attempt to analyze the spatial-temporal distribution of the taxi trips data of New York City (NYC) throughout 2014. Our primary goal is to find the vibrant neighborhoods favored by New Yorkers during night time. By adopting spectral clustering, we successfully identify the "day-time" neighborhoods and the "night-time" neighborhoods in NYC, which is reasonably in accordance with the reality.**

***Index Terms*—Unsupervised learning, spectral clustering, NYC taxi trips.**

I. Introduction

WITH the availability of a wide range of public civic data sets and open source tools for data science, nowadays anyone with a fundamental knowledge/experience on statistics and programming is well-equipped to "learn" about a city in various aspects, a realm used to be dominated primarily by professional data scientists and/or urban planners. In this work, we consider the general problem of identifying the "active" neighborhoods in NYC during the night time. An answer to it can be made use of, for example, by tourists seeking for exciting nightlife, new residents looking for a quiet neighborhood, taxi drivers to locate prospective passengers, or city planners to predict traffic, to name a few.

A simple approach to answer this question qualitatively, of course, is to ask the locals or search for an answer on Google or TripAdvisor, etc. However, to get a more quantative, founded and potentially more useful answer, we take a rigorous approach to look for it from the (NYC) Taxi & Limousine Commission (TLC) Trip Record Data [1] during 2014 and the 2010 NYC Census Tracts (CTs) maps [2]. The number of taxi pickup and dropoffs per hour are used as an indicator on how active each CT is. Most analysis on similar datasets are based on supervised learning, such as the MIT 2013-2014 Big Data Challenge [3], a related course project [4], as well as the recent algorithm competition by DiDi Research [5]. On the contrary, our analysis is primarily based on a simple but effective unsupervised learning technique, namely spectral clustering. With straightforward visualization results, we successfully reveal the distinct "day-time" and "night-time" neighborhoods of NYC.

The rest of this report is organized as follows. Section II introduces the data sets and techniques we used to clean and import the data. Section III describes how we adopt the spectral-clustering method to cluster the CTs into two groups according to the pickup-dropoff patterns. Numerical results are provided in Section IV, which clearly shows that the two distinct clusters of the busiest CTs indeed represent the neighborhoods more active during daytime/nighttime. Finally, Section V concludes the report.

W. Wu is with the Department of Electrical and Computer Engineering, University of California, Davis, CA 95616 USA (e-mail: wnhwu@ucdavis.edu).

II. The Data Sets

The primary dataset studied in this project is the NYC TLC Trip Record Data [1] during 2014 including both the yellow and green taxi trips, which in csv formats amount to a total of around 25GB. Among various fields, we are mainly interested in the pickup/dropoff time and location of each trip. Also, as the original data consists of trips of both yellow and green taxi, which have different modus operandi and coverage [6], each record of trip is also labeled as "yellow" or "green".

In order to analyze the taxi trips at a proper spatial granularity, we adopt the 2010 NYC census tracts map [2] which divides the city into 2166 CTs as in [7], each composed of around 5-10 blocks. To enable efficient manipulations of the data, especially mapping the longitude and latitude of each pickup and dropoff to a CT, the trip data are firstly written into a PostgreSQL database with PostGIS extension using Python package `sqlalchemy` with `geoalchemy2`. After data cleansing, a total of 176493349 records are inserted into the database. The CT map in shapefile format is added to the database directly with `shp2pgsql` data loader.

III. Distinguish Neighborhoods Vibrant in the Night

A. Spatial-Temporal Distribution of Taxi Trips

We count the number of pickups and dropoffs for each hour of the day (0-23) for each CT (1-2166) averaged over the 365 days of 2014. The results are represented as a 2166-by-48 matrix $\mathbf{C}$, where $C_{i,j+1}$ and $C_{i,25+j}$, $i \in [1, 2166]$,, $j \in [0, 23]$ represent the average number of pickups/dropoffs in the i-th CT from j to $j+1$, respectively. Denote the 2166-by-1 vector $\mathbf{s}$ as the area of each CT, then the average density of pickups/dropoffs, namely $\text{diag}(\mathbf{s})\mathbf{C}$, for yellow and green taxis are shown in Fig. 1, Fig. 2, Fig. 3 and Fig. 4, respectively.

In order to explore the temporal pattern of the taxi trips, each row of $\mathbf{C}$ is normalized to have sum of 1, resulting in $\bar{\mathbf{C}} = [\text{diag}(\mathbf{C}\mathbf{1})]^{-1}\mathbf{C}$, where $\mathbf{1}$ is a 48-by-1 all-one vector. Also as shown in the density maps, for both yellow and green taxi most pickups and dropoffs are highly concentrated in a few CTs. Consequently, we focus only on the most popularized CTs, i.e. we select a subset of rows from $\bar{\mathbf{C}}$ as $\bar{\mathbf{C}}_S$. For yellow taxi, we select all $N_S = 81$ CTs with average daily total pickups/dropoffs greater than 4500. For green taxi, we select all $N_S = 44$ CTs with average daily total pickups/dropoffs greater than 300.

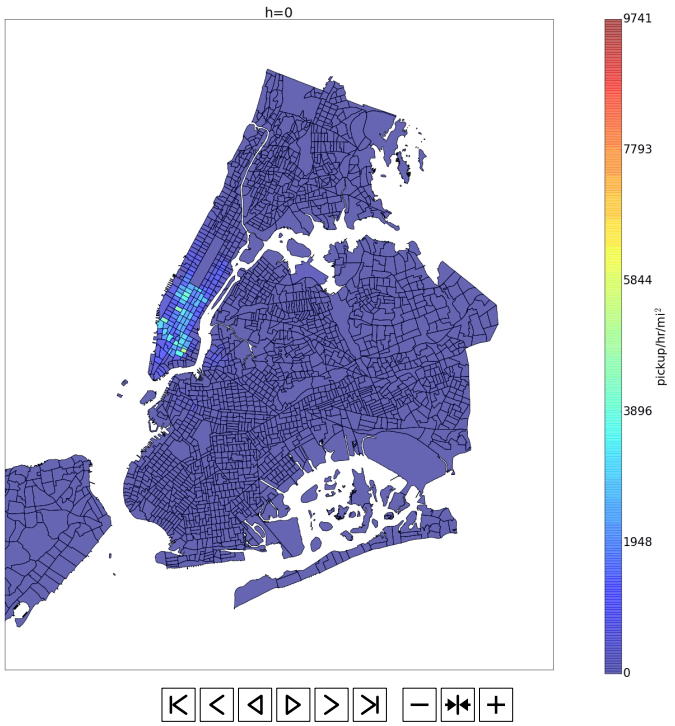


Fig. 1. Density of pickups, yellow taxi.

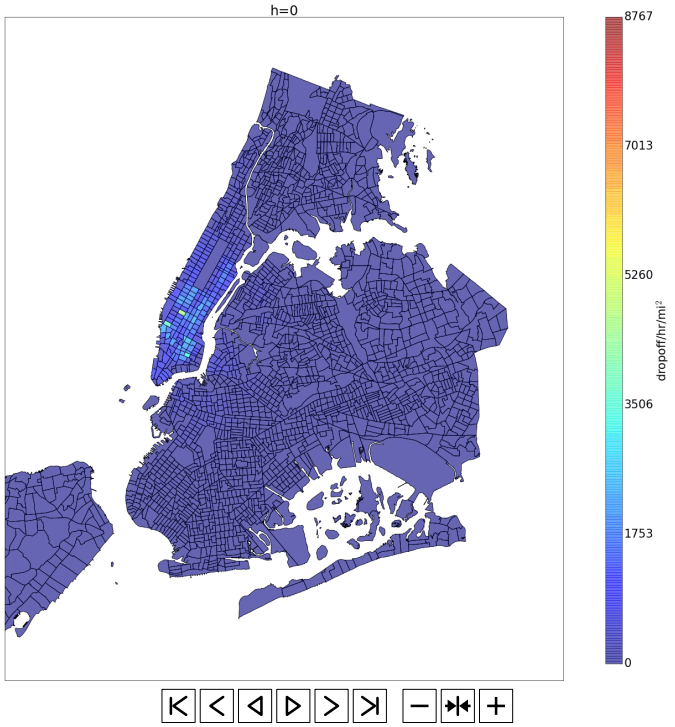


Fig. 2. Density of dropoffs, yellow taxi.

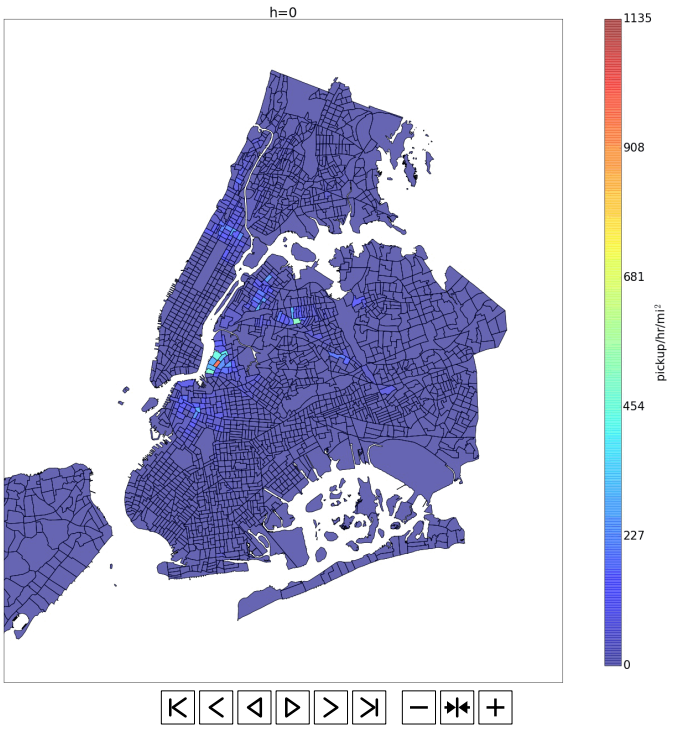


Fig. 3. Density of pickups, green taxi.

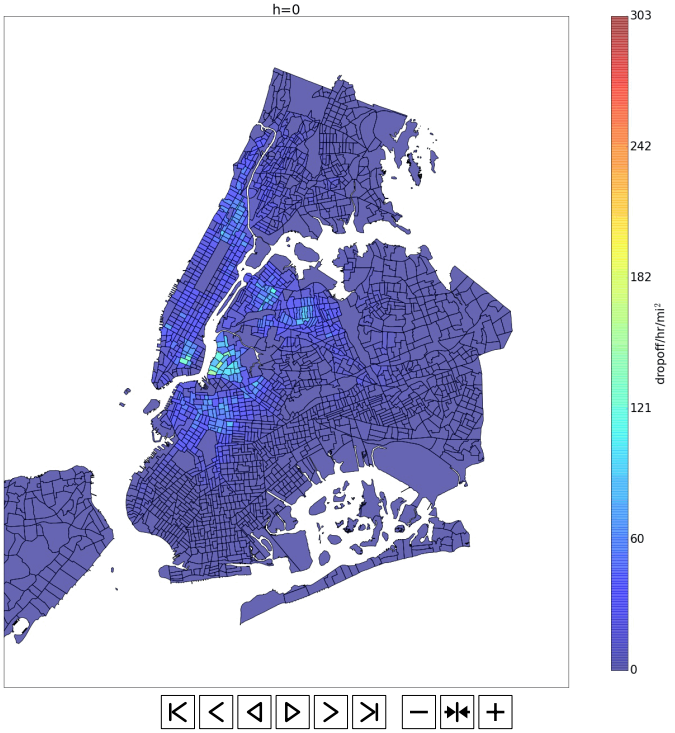


Fig. 4. Density of dropoffs, green taxi.

B. Spectral Clustering

To group the CTs based on their daily pickup/dropoff patterns, we perform spectral clustering [8] on $\bar{\mathbf{C}}_S$ for yellow and green taxi, respectively. Specifically, we evaluate the N_S-by-N_S Gaussian radial basis function (rbf) distance matrix $\mathbf{W}$ where

$$W_{i,j} = \exp\left(\frac{\|\mathbf{c}_i - \mathbf{c}_j\|_2^2}{\epsilon}\right) \tag{1}$$

where $\mathbf{c}_i$, $\mathbf{c}_j$ are the i, j-th row of $\bar{\mathbf{C}}_S$. Then we apply k-means on $\mathbf{v}$, namely the second principle eigen-vector of $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where the diagonal degree matrix $\mathbf{D} = \text{diag}(\mathbf{W1})$.

The parameter ϵ is selected by examining the historgram of entries of $\mathbf{v}$. In our case, we select ϵ to be 180 times the median of $\|\mathbf{c}_i - \mathbf{c}_j\|_2^2$ for yellow taxi trips and 160 times the median for green taxi trips.

IV. Numerical Results

The histogram of the entries of $\mathbf{v}$ for yellow taxi and green taxi are shown in Fig. 5 and Fig. 6, respectively. The spectral clustering indeed identifies two clusters, and the corresponding CTs are plotted in Fig. 7 and Fig. 8 with Python's `matplotlib` library with `Basemap` toolkits. Cluster 0 primarily includes upper and middle Manhattan, JFK Airport, LaGuardia Airport, and the north-west part of Queens and Brooklyn. Cluster 1, on the other hand, primarily includes certain areas of West Village, East Village and Williamsburg. By examining the temporal variation of pickups and dropoffs for the two clusters as shown in Fig. 9 and Fig. 10, the two clusters are different in the following aspects:

- Cluster 0 has more pickups/dropoffs during the day time (5:00-17:00) than cluster 1.
- Both the pickup/dropoff curve reach their peak at around 18:00/19:00 for cluster 0. In comparison, for cluster 1, the pickup curve reaches its peak at around 0:00, while the drop off curve reaches its peak at around 20:00.

Consequently, it is reasonable to argue that cluster 0 represents the most popularized "day-time" neighborhoods of NYC while cluster 1 represents the most popularized "night-time" neighborhoods. This is also in accordance with the results shown in the "NYC Late Night Taxi Index" section of [7]. To further justify the clustering results, we plot the top 20 most reviewed bars/lounges, dance clubs and formal restaurants from Yelp in Fig. 11. The "night-time" cluster of 13 CTs found by the spectral clustering algorithm contains around 1/3 of the 60 most reviewed hotspots, while the business area of midtown-midtown south and the Financial districts are correctly (heuristically) classified as the "day-time" neighborhoods. However, we admit that the spectral clustering method adopted here is rather sensitive to the choice of parameter N_S and ϵ, and the interpretation of the results is subjective as in most unsupervised learning techniques. More rigorous approaches to similar problems can be found in [9] and the references therein.

V. Conclusion

In this work, we studied the problem of infering the nightlife hotspot in New York City by making use of the NYC taxi data in 2014. By extracting the daily pickup/dropoff patterns sampled hourly and adopting the spectral clustering algorithm on the high dimensional (48) feature, we are able to reasonably cluster the CTs into the "day-time" and "night-time" neighborhoods.

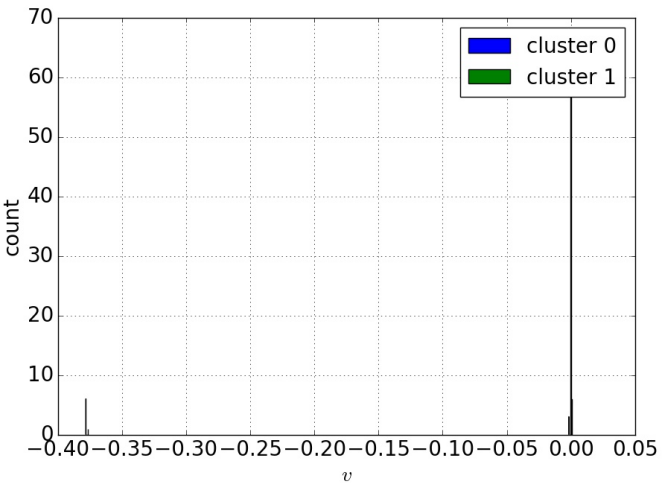


Fig. 5. Histogram of $\mathbf{v}$ for yellow taxi.

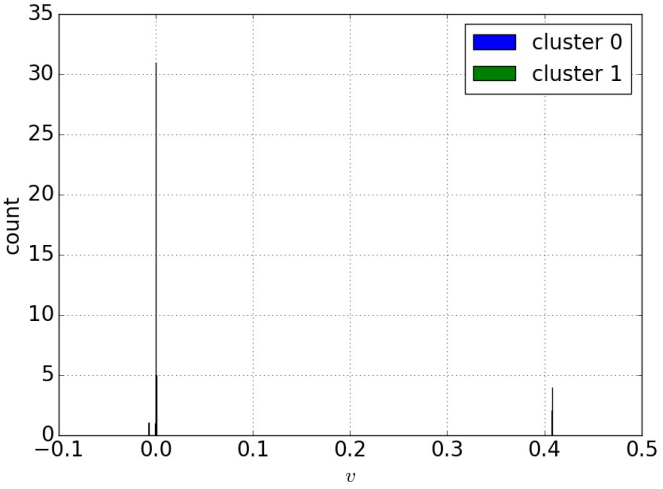


Fig. 6. Histogram of $\mathbf{v}$ for green taxi.

References

[1] N. T. . L. Commission. (2016) NYC TLC Trip Data. [Online]. Available: http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

[2] N. Department of City Planning. (2014) Census Tracts. [Online]. Available: https://data.cityofnewyork.us/City-Government/2010-Census-Tracts/fxpq-c8ku

[3] MIT Computer Science and Artificial Intelligence Lab (CSAIL). MIT Big Data Challenge. [Online]. Available: http://bigdata.csail.mit.edu/challenge

[4] J. Grinberg, A. Jain, and V. Choksi, "Predicting taxi pickups in new york city," *Final Paper for CS221 Artificial Intelligence, Computer Science Department, Stanford University*, 2014.

[5] Didi Research Institute. Di-Tech Challenge. [Online]. Available: http://research.xiaojukeji.com/competition/

[6] Quora. What is the difference between Green Cabs and Yellow Cabs? [Online]. Available: https://www.quora.com/What-is-the-difference-between-Green-Cabs-and-Yellow-Cabs

[7] T. W. Schneider. (2015) Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance. [Online]. Available: http://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/

[8] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[9] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12. New York, NY, USA: ACM, 2012, pp. 186–194. [Online]. Available: http://doi.acm.org/10.1145/2339530.2339561

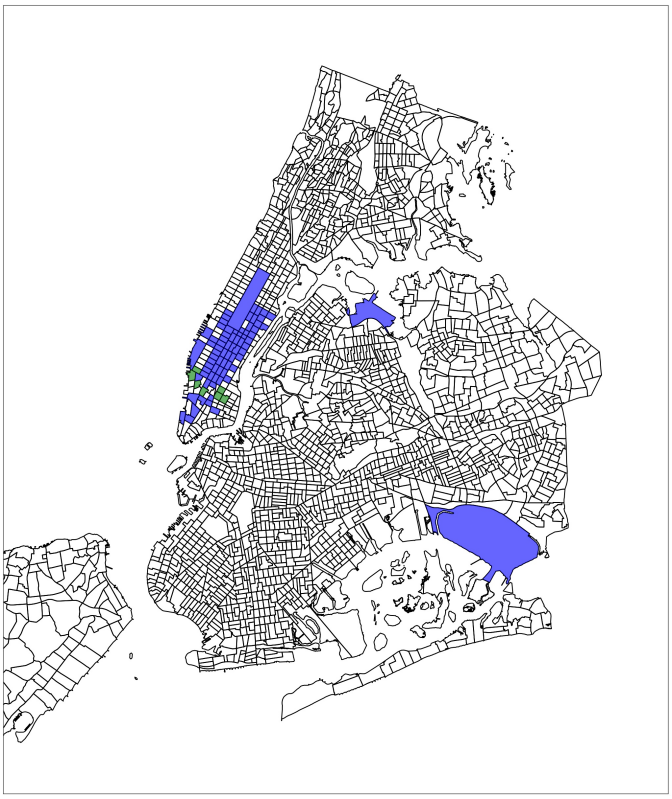


Fig. 7. Spatial distribution of the two cluster of CTs for yellow taxi.

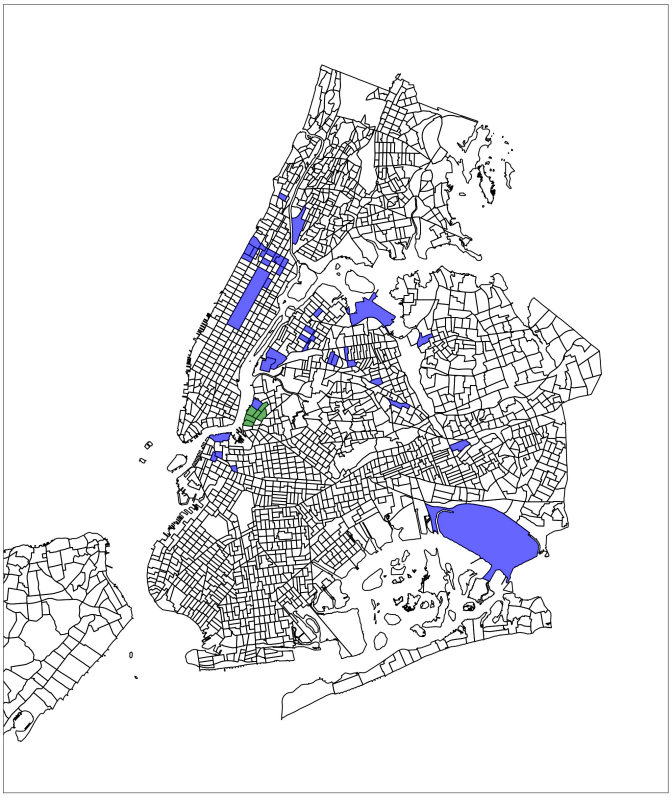


Fig. 8. Spatial distribution of the two cluster of CTs for green taxi.

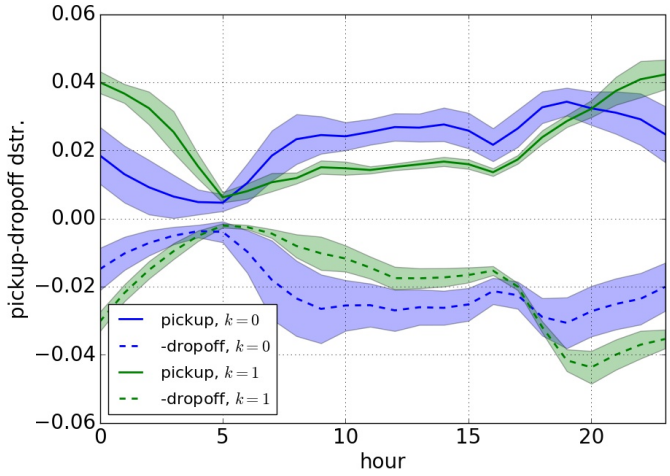


Fig. 9. Temporal variation of pickups/dropoffs of the two clusters of CTs for yellow taxi. Lines represent the mean value and shaded areas represent the standard deviation.

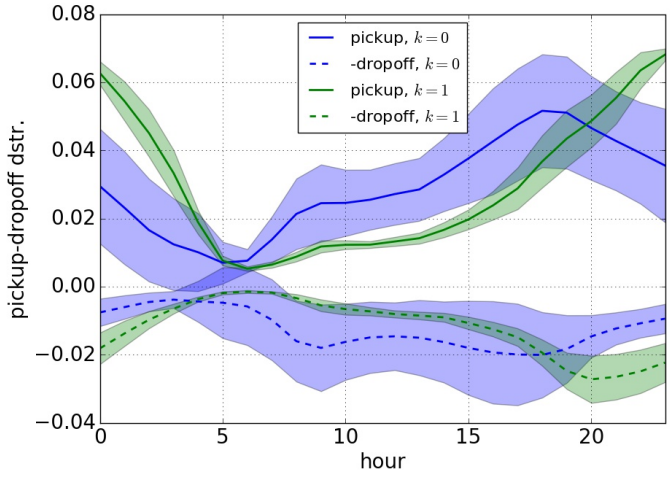
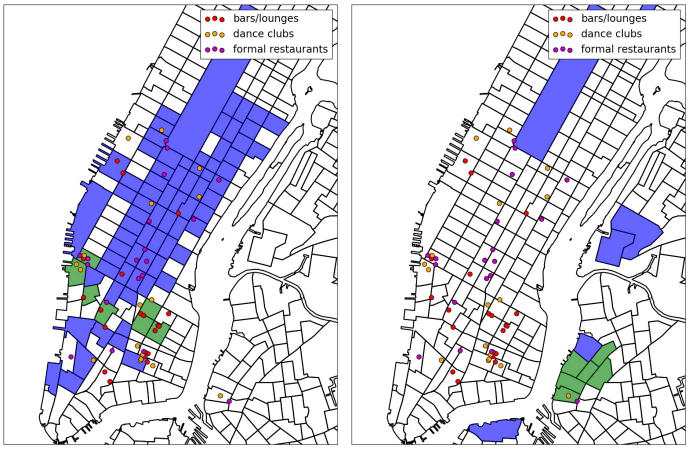


Fig. 10. Temporal variation of pickups/dropoffs of the two cluster of CTs for green taxi.



(a) Yellow (b) Green

Fig. 11. Spatial distribution of the two cluster of CTs plotted with the top 20 most reviewed bars/lounges, dance clubs and formal restaurants from Yelp.