Exploring demand patterns of a ride-sourcing service using spatial and temporal clustering

Liu, T. L.K.; Krishnakumari, P.; Cats, O.

DOI
10.1109/MTITS.2019.8883312

Publication date
2019

Document Version
Final published version

Published in
MT-ITS 2019 - 6th International Conference on Models and Technologies for Intelligent Transportation Systems

Exploring Demand Patterns of a Ride-Sourcing Service using Spatial and Temporal Clustering

T.L.K. Liu
Dept. of Transport and Planning
Delft University of Technology
Delft, The Netherlands
t.l.k.liu@student.tudelft.nl

P. Krishnakumari
Dept. of Transport and Planning
Delft University of Technology
Delft, The Netherlands

O. Cats
Dept. of Transport and Planning
Delft University of Technology
Delft, The Netherlands

***Abstract*—On-demand transport has become a common mode of transport with ride-sourcing companies like Uber, Lyft and Didi transforming the mobility market. Recurrent patterns in prevailing demand patterns can be used by service providers to better anticipate future demand distribution and thus support demand-anticipatory fleet management strategies. To this end, we propose three steps for extracting such demand patterns from travel requests: (1) constructing the origin-destination zones by spatial clustering, (2) composing the hourly and daily origin-destination matrix, and; (3) temporal clustering to extract the dynamic demand patterns. We demonstrate the three step approach on the open-source Didi ride-sourcing data. The data consists of travel requests data for November 2016 from Chengdu, China amounting to approximately 6 million rides. The analysis reveals pronounced and recurrent and thus predictable daily and weekly patterns with distinct spatial properties pertaining to ride-sourcing production and attraction characteristics.**

***Index Terms*—ride-sourcing, spatial clustering, temporal clustering, demand patterns, taxi data**

I. Introduction

Technical developments of smartphones integrating GPS functionality, internet connectivity and trust in online marketplaces makes it possible for ride-sourcing to evolve to phase five, a technology-enabled ride-matching [1]. These ride-sourcing companies (also known as TNCs, Transportation Network Companies in the US) are creating online marketplace platforms that allow matching incoming travel requests and registered drivers in real-time. The estimated worldwide market value of ride-sourcing services in May 2018 was over 150 billion U.S. Dollars with Uber, Didi Chuxing (Didi) and Lyft on top with respectively a valuation of 72, 56 and 11.5 billion U.S. dollars [2]. Didi is the largest ride-sourcing company in terms of operations with 30 million rides per day and 21 million drivers, twice and seven times as much as Uber, respectively [3]–[5].

Given the recent rapid growth of ride-sourcing services and the commercial sensitivity, there is a great interest yet only very limited knowledge insofar on their demand patterns. Notwithstanding, recent studies examined the impacts of ride-sourcing on the taxi market [6], passengers' response to dynamic pricing [7]. In addition, there are also pioneering efforts to develop methods for short-term demand predictions for ride-sourcing services using deep-learning approaches that utilize both spatial and temporal relations [8], [9], as well as proposing a framework for predicting overall system performance [10]. These studies provide initial insights into the potential to mine historical demand data for improved forecasts. All of these studies have adopted an artificial intelligence approach and employed machine learning techniques for investigating ride-sourcing data.

The approach taken in this study is anchored in transport demand analysis and is aimed at identifying meaningful spatial and temporal clusters in demand for ride-sourcing services to better understand the underlying patterns and support planners as well as service providers. The latter can better cater for prevailing demand patterns by deploying proactive pricing and fleet management strategies in anticipation of recurrent demand characteristics and thus contribute to the efficiency and effectiveness of the service provisioned. Previous research has demonstrated that demand-anticipatory dispatching algorithms that leverage on demand predictions can reduce passengers waiting and in-vehicle times [11].

Clustering allows to reduce the computational resources and dimensionality of the data which contains millions of disaggregate origin and destination locations. Furthermore, we assess the spatial features of the demand profiles in terms of urban areas with a surplus or deficit (i.e. more rides are attracted than generated to a given zone or vice-versa, respectively) of rides within various temporal profiles. We perform our analysis for a dataset of Didi services in Chengdu, China.

The remainder of this paper is structured as follows. Section II describes the methodology. Section III details the experimental setup. The results of the spatial clustering and the temporal clustering are discussed in section IV. Finally, section V concludes with the key findings, limitations, and future research.

II. Methodology

In this paper, we propose three steps for extracting demand patterns from travel requests - (1) constructing the origin-destination zones by spatial clustering (2) calculating time-dependent origin-destination matrix, and (3) temporal cluster-

978-1-5386-9484-8/19/$31.00 ©2019 IEEE

ing to extract the dynamic demand patterns. The flowchart (Fig. 1) shows an overview of the inputs and outputs for the three steps.

The clustering step for creating static OD zones is the spatial clustering. Once the static zones are defined, we can create the OD matrices at different aggregation levels which tells us the number of trips between each origin zone to destination zone at different levels. Such OD matrix can compactly define the ride-sourcing demand for the entire network. Finally, the OD matrices at different levels can be investigated to check if distinct demand patterns emerges in the city which can, ultimately, make the demand predictable. Fig. 1 shows an overview of the framework.

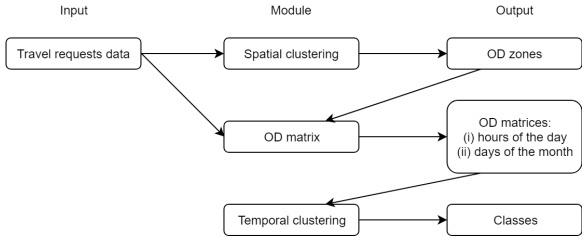


Fig. 1: Flow chart of analysis framework

A. Spatial clustering

Travel request data is geo-located with an origin (pick-up) location and destination (drop-off) location with given time-stamps for boarding and disembarking for each ride. There are millions of such unique geo-locations in the travel request data. Hence, the first step for understanding the demand patterns is to reduce the dimension of the locations by grouping/clustering together origin-destination locations to create meaningful and compact origin-destination zones. In this work, we use the well-known k-means algorithm described in [12] for the clustering.

Given a set of d-dimensional geo-locations $(x_1, x_2, ...x_n)$, the k-means clustering aims at grouping the n observations into k zones $(Z_1, Z_2, ...Z_k)$ that minimizes certain criteria. The minimization criteria is the within-cluster distance between the points in the zone i to the its cluster centroid c_i for a given k defined as:

$$argmin \sum_{i=1}^{k} \sum_{x \in Z_i} ||x - c_i||^2 \tag{1}$$

k centroids are randomly selected from the geo-locations and then they are iteratively updated N times until the minimum criteria given in (1) is reached for a given k. These centroids are considered optimal for that k. One of the main disadvantage of using k-means or any unsupervised clustering is that the number of clusters k needs to be supplied as a parameter. In this study, we use four metrics proposed in [13] to quantitatively compute the optimal k to ensure creating compact and meaningful clusters. These are two distance metrics - intra-distance d_{intra} and inter-distance d_{inter}; and two flow metrics - intra-flow q_{intra} and inter-flow q_{inter}.

Compact clusters can be created by using the distance metrics by minimizing the distance between the cluster centroids and the points in the cluster (d_{intra}) as formulated in (2) while maximizing the distance between centroids of different clusters (d_{inter}) as defined in (3).

$$d^{intra} = \sum_{i=1}^{k} \sum_{x \in Z_i} d(x, c_i) \tag{2}$$

where $d(x, c_i)$ is the geodesic distance between a geo-location x in zone Z_i and the centroid c_i of the same zone.

$$d^{inter} = \sum_{i=1}^{k} \sum_{j=1}^{k} d(c_i, c_j) \tag{3}$$

where $d(c_i, c_j)$ is the geodesic distance between centroid c_i of zone Z_i and centroid c_j of zone Z_j.

Another important metric considered for finding the optimal k is the flow metric. In this work, instantaneous flow between geo-locations x and y is defined as $q(x, y, t, \tau)$ and is the number of trips that originates at location x with destination at location y at time t for a given day τ. This can be derived directly from the travel request data. Since the spatial clustering is for creating static zones, we only need to know the total flow between different the k clusters and not the instantaneous flow between geo-locations. The total flow between zone Z_i and Z_j is defined as:

$$Q(i, j) = \sum_{x \in Z_i} \sum_{y \in Z_j} \sum_{t=1}^{T} \sum_{\tau=1}^{D} q(x, y, t, \tau) \tag{4}$$

where t ranges from $[1, T]$ corresponding to time $[00:00, 24:00]$ and D is the total number of days available in the travel request dataset. We consider two flow metrics - inter-cluster flow q^{inter} and intra-cluster flow q^{intra}. The aim of using the flow metric is to find optimal k that maximizes the inter-cluster flow as defined in (6) and minimises the intra-cluster flow as defined in (5) so as to create distinct clusters that have high flow between the different zones than high flow just within the zones.

$$q^{intra} = \frac{\sum_{i=1}^{k} Q(i, i)}{\sum_{i=1}^{k} \sum_{j=1}^{k} Q(i, j)} \tag{5}$$

where $Q(i, i)$ is the flow that originates in zone Z_i with destination in zone Z_i as well and $Q(i, j)$ is the flow with origin at zone Z_i and destination at zone Z_j.

$$q^{inter} = \frac{\sum_{i \neq j} Q(i, j)}{\sum_{i=1}^{k} \sum_{j=1}^{k} Q(i, j)} \tag{6}$$

These four metrics related to spatial distance and flow are computed for different k and then they are investigated to select the final number of spatial zones k^* for the dataset. The entire process for estimating the k^* is given in Algorithm 1.

Algorithm 1: Spatial Clustering

Function
- **Input :** geo-locations x, number of zones k, number of random centroids R, iterations N
- **Output:** Centroids c_k, labelled geo-locations x'
- **foreach** *K = 1 to k* **do**
  - `/* Perform k-means clustering */`
  - **foreach** *r = 1 to R* **do**
    - **foreach** *iter = 1 to N* **do**
      - $c^K_{iter,r} \leftarrow K$ random centroids from x
      - $x'_{iter,r} \leftarrow$ map x into these K centroids
      - $\text{error}_{iter,r} \leftarrow$ Eq:(1) for $k = K$
  - $x'_K \leftarrow x'_{iter,r},\ c_K \leftarrow c^K_{iter,r}$ that minimizes $\text{error}_{iter,r}$
  - `/* Compute distance metric */`
  - Compute d^{intra}_K, d^{inter}_K
  - `/* Compute flow metric */`
  - Compute q^{intra}_K, q^{inter}_K
  - $x' \leftarrow x'_K,\ c_k \leftarrow C_K$ that $min(d^{intra}_K, q^{intra}_K)$ & $max(d^{inter}_K, q^{inter}_K)$

B. OD matrix computation

With these static zones as the origins and destinations, we can compute the OD matrix to represent the demand of the network. Each cell in the OD matrix corresponds to the number of trips started within a time period from a particular origin to a particular destination for a given day. The OD matrix can be used to understand where the demand is produced and attracted with respect to the zones. Depending on the application, the OD matrix can be computed at different aggregation levels. In this work, we aggregate the OD matrix at different levels - hourly and daily. The hourly OD matrix $Q(i, j, t)$ for all the days in the dataset at time t which ranges from [00:00,24:00] in increments of one hour is defined as:

$$Q(i,j,t) = \sum_{x \in Z_i} \sum_{y \in Z_j} \sum_{\tau} q(x,y,t,\tau) \tag{7}$$

Thus, there are 24 $k \times k$ hourly OD matrices for the whole dataset, where k is the number of spatial zones.

The daily OD matrix $Q(i, j, \tau)$ for day τ with time ranging from [00:00,24:00] is defined as:

$$Q(i,j,\tau) = \sum_{x \in Z_i} \sum_{y \in Z_j} \sum_{t} q(x,y,t,\tau) \tag{8}$$

where τ determines the day. Thus, there are D $k \times k$ daily OD matrices for the whole dataset where D is the number of days that have travel request data available.

C. Temporal clustering

For studying the demand dynamics of the spatial zones, we use the vectorized form of the hourly and daily aggregated OD matrices as feature vectors for the temporal clustering. The aim of this is to study if regular patterns emerge for different time periods of a day or different days. These insights can be used for developing demand-oriented fleet management for different time periods within a day and between days. In this work, we use hierarchical agglomerative clustering for the temporal clustering [14]. This is because of the power of such clustering in revealing the distribution of the feature vectors in the form of a dendrogram which can aid in determining the optimal number of clusters as shown in [15]. Dendrogram is a tree diagram illustrating the arrangement of the clusters [16]. The hierarchy or distribution of the feature vectors is constructed based on a dissimilarity metric between the feature vectors. The dissimilarity metric used in this work is the city block distance and the connectivity between any two d-dimensional feature vectors, u, v is determined based on the city block distance as:

$$d(u,v) = \sum_{i=1}^{d} |u_i - v_i| \tag{9}$$

The temporal clustering is performed on both the hourly and daily OD matrices separately. The $k \times k$ OD matrix is vectorized to obtain one dimensional feature vector of dimension $k * k$ for each hour and each day respectively. Each feature vector is initialised as a single cluster and then two clusters will be merged based on the dissimilarity metric and this merging process continues until only one cluster remains. The result of this merging can be illustrated using the dendrogram which can be used to decide the number of clusters.

For each cluster, we build a representative OD matrix inorder to make the analysis informative and comprehensible. The representative OD matrix of the cluster is the medoid of all the OD matrices that belongs to the clusters defined as:

$$x_{medoid} = \underset{y \in \{x_1, x_2, \ldots, x_n\}}{\arg\min} \sum_{i=1}^{n} d(y, x_i) \tag{10}$$

where $x_1, x_2, ...x_n$ are a set of n feature vectors in a given cluster with d-dimensional real vectors and $d(y, x_i)$ is the pair-wise dissimilarity metric. A mediod is an object in the cluster that minimizes the dissimilarity to all other objects in that cluster. We considered the medoid instead of the centroid as the medoid is an actual data point corresponding to an actual demand.

III. Experimental Setup

We demonstrate the three step approach on the open-source Didi taxi data for the region of Chengdu, China. In this section, we explain the Didi data, some descriptive statistics of the data and define the parameter choices for the clustering and some of the key metrics used to explain the clustering.

A. Data

The open-sourced Didi data is composed of 1 month (November 2016) of travel requests data from a small area in Chengdu, China, with approximately 200 000 rides for a single day on average. Chengdu is a city that has been gaining more economical importance over the years as it has

been rapidly developing and becoming a main hub for several different industries [17]. Like many Chinese cities, Chengdu has a large urban area of more than 1,700 km^2. The capital of the Sichuan province currently has a population of almost 11.5 million inhabitants, and a population density of 6,500 people per km^2 [18].

The travel request data obtained from the GAIA initative of Didi Chuxing [19] is used in this paper. The ride request data is available for all rides started in November 2016 which have at least one Global Positioning System (GPS) point within a specified area of Chengdu as shown in Figure 2. Some of the origin and destination points in the travel request data are outside this specified area. However, the travel requests outside the specified GPS trace area is not complete, hence there is a inherent bias in the data which might be reflected in the demand patterns as well. The travel request data includes the origin point (pick-up location), destination point (drop-off location), ride start time and ride end time. The trip route data which contains the GPS traces are included separately, however that data is not used in this study.

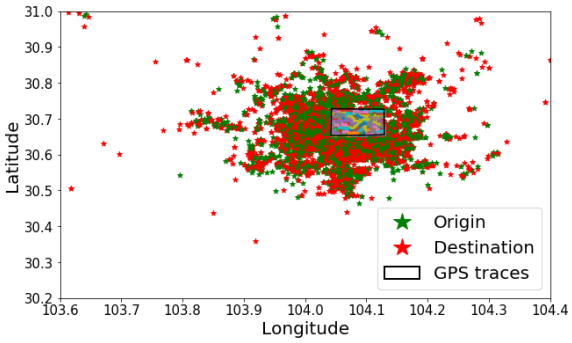


Fig. 2: Spatial plot of the first 25000 rides from the DiDi data with origins, destinations and GPS traces. The real size of the plot is about 76.7 x 88.7 km (width x height) and the rectangle with the GPS traces is about 8.3 x 8.1 km (width x height).

The dataset is cleaned by removing the orders that are repeated with the same order ID, which is approximately 15% of the requests. Our assumption is that the duplicated requests correspond to multiple passengers per ride or ride-sharing and these are not considered within this study. Also, some outliers (126 rides) with geodesic trip distance higher than 400 km are removed. The cleaned dataset contains 6,104,877 unique rides for 30 days of November 2016.

B. Descriptive statistics

For understanding the dynamics of the specified area, the following descriptive statistics are considered of the travel requests:

- temporal distribution over the day,
- geodesic distance of the trip, and
- trip travel time.

From Figure 3, it can be seen that the temporal distribution of the travel request data over the day binned with different time periods is relatively stable between 09:00-21:00. In comparison to traditional modalities, there is no distinct morning and evening peaks. The morning peak peaks between 09:00-09:30, which is quite late in comparison to traditional modalities. Here, those traditional rush hour peaks are less pronounced, while the highest peak is between 13:30 to 14:00 implying an afternoon peak.

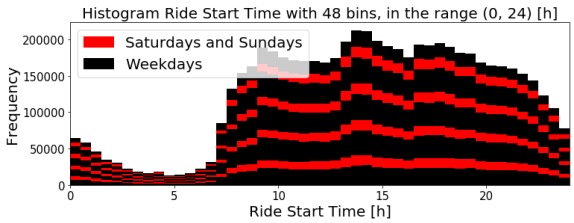


Fig. 3: Distribution of the travel requests over different time periods, in the range (0, 24) [h]

Figure 4 shows the distribution of geodesic distances of each travel request and it shows that the service is mainly used for rides shorter than 10 km. The peak is between 3-4 km. The average geodesic distance is 6.44 km and the median is 5.28 km.

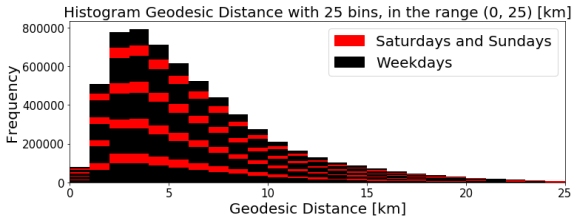


Fig. 4: Distribution of the travel requests based on the geeodesic distance of the each trip, in the range (0, 25) [km]

The histogram of travel times in Figure 5 shows that most rides are between 5-35 minutes. The peak is between 10-20 minutes. The average ride duration is 22.12 minutes and the median is 19.28 minutes.

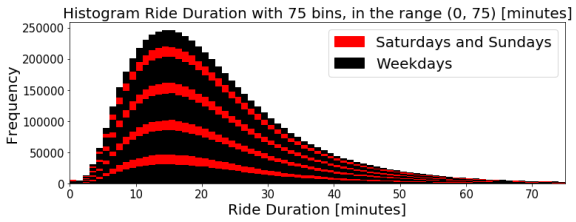


Fig. 5: Distribution of the travel requests based on the ride duration, in the range (0, 75) [minutes]

The distance and travel time distributions follow a positive skewed bell shape, meaning the mass is in the shorter rides. Thus, it can be said that Didi is mainly be used for short rides with an average geodesic distance of 6.44km and an average ride duration of 22.12 minutes.

C. Parameters choices

There is a couple of parameters that needs to be set for the different algorithms involved in the three step approach. First, for the spatial clustering, we ran the algorithm for different number of clusters ranging from 5 to 100. An optimal number of clusters is chosen based on the different defined metrics within these range of clusters. For the k-means itself, there are two key parameters - the number of initial centroids which is set to 10 in this work and maximum number of iterations is set to 300. The maximum number of iterations is an additional stopping convergence criteria for the k-means clustering.

For the temporal clustering, we set mainly two parameters for the hierarchical agglomerative clustering - the connectivity method and the distance metric. The connectivity method used is average and the distance metric is the city block distance which uses the sum of the absolute difference between all OD-pairs of different OD-matrices. The optimal number of clusters is chosen based on the insights gained from the dendrogram.

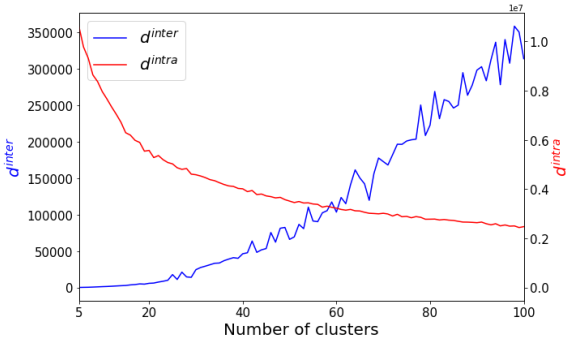
IV. Results

In this section, we report the results of the spatial clustering and analyse the resulting zones and their corresponding descriptive statistics. The result of the temporal clustering of different time periods and days are illustrated using the tree diagram - dendrogram and the resulting clusters are analysed to infer the type of demand each of these clusters represents.

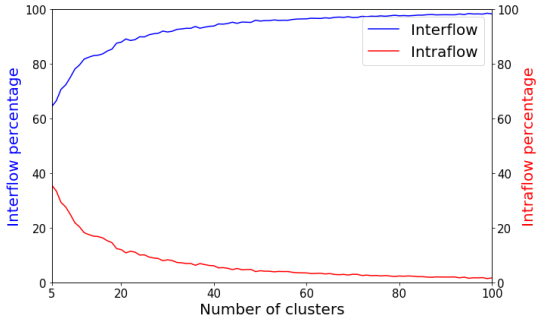
A. Spatial clustering

We performed a sensitivity of number of clusters on the spatial clustering based on four metrics - two distance metrics (d^{intra}, d^{inter}) and two flow metrics (q^{intra}, q^{inter}). The results of these four metrics for varying number of clusters are shown in Figure 6. Both intra distance and intra flow is a decreasing function with the first increments dropping rapidly with the number of clusters and then decreasing at a slower pace for larger number of clusters. This was to be expected as the number of clusters increases, each cluster is composed of less elements but with less dissimilarity in terms of distance and flow while the total flow remains constant. The property for the inter distance and inter flow is an increasing function with the dissimilarity between the clusters increasing with higher number of clusters as shown in Figure 6.

The optimal number of clusters would ideally have large dissimilarity between clusters (high inter distance and inter flow) and small dissimilarity within clusters (low intra distance and intra flow). This implies a larger number of clusters as evident from the continously increasing and decreasing functions of the metrics. However, with large number of clusters, the clusters become less interpretable and more complex. Hence, there needs to be a trade-off between complexity and optimization. Based on these metrics and considering the trade-off, the number of clusters is chosen as 50 as they are a comprehensible number of zones and after around 40 clusters, the rate at which the metrics increases or decreases have started becoming more stable. This results in the 50 spatial zones as shown in Figure 7.



(a)



(b)

Fig. 6: (a) Intra and inter distance (b) Intra and inter flow for 5 to 100 clusters. Each clustering is achieved with 300 iterations and 10 random centroid seeds.

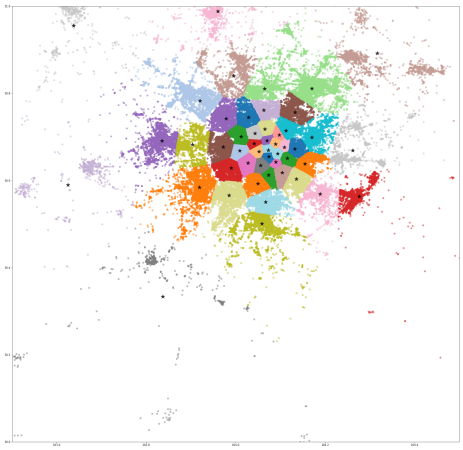
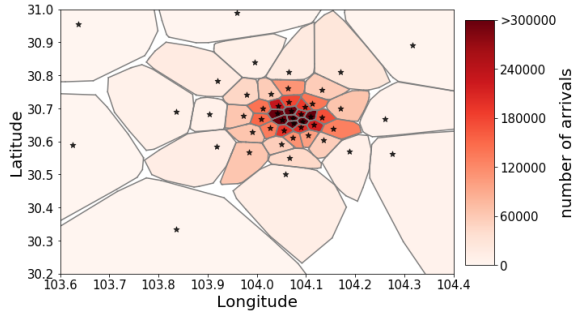


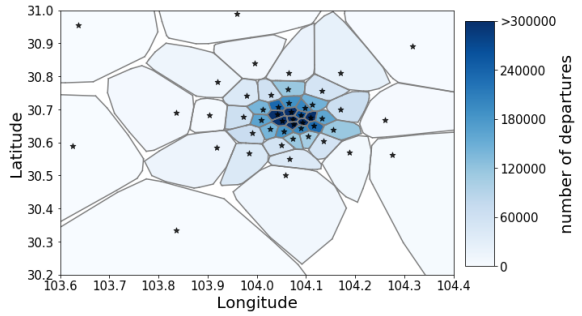
Fig. 7: 50 zones obtained after the spatial clustering

A look at the number of rides for different zones shows that there is around 500,000 rides per month in the center whereas the number of rides in the outer zones are in the range of 100,000 rides or less. A breakdown of these rides into rides that originates from a zone (production) or have it as the ride destination (attraction) is shown in Figure 8. There is large

variation between the center and the outer zones indicated by the large color differences between the different zones. The center zones with the small spatial dimension has a much higher number of rides in comparison to the neighbouring zones.



(a)



(b)

Fig. 8: (a) Number of arrivals per zone or Attraction (b) Number of departures per zone or Production; for the month of November 2016.

This is also clearly evident from the cumulative density functions (CDF) of the production and attraction given in Figure 9. Both production and attraction shows similar cdf plots. The differences between the zones with less rides and zones with more rides is large. 20% of the zones have less than 10,000 departures or arrivals in the month of November while the top 20% has more than 200,000 departures or arrivals.

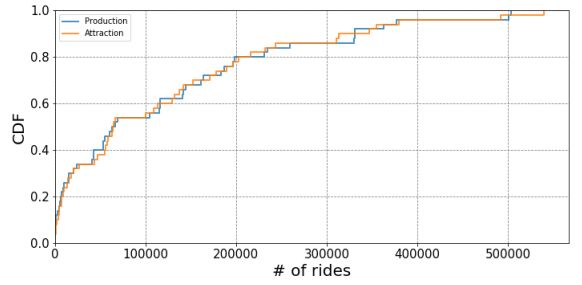


Fig. 9: CDF of production and attraction of the 50 zones

From the Figure 8, the production and attraction volumes seems proportionally correlated with each other in all respective zones. However, a closer look at the actual differences between productions and attractions in Figure 10 shows more spatial variability. This is further evident from Figure 11 which shows the relative share of arrivals in the different zones, i.e, number of arrivals divided by the total number of arrivals and departures in that zone (total flow). High value implies a surplus of arrivals while a low value implies deficit.

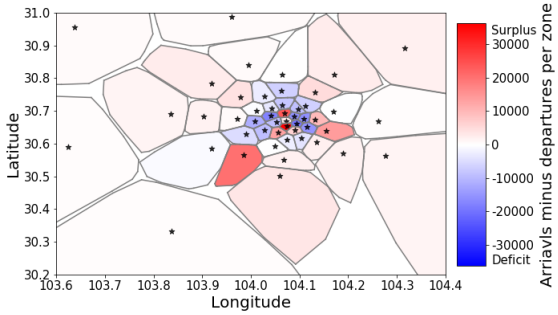


Fig. 10: Difference between number of arrivals and departures per zone.

From the Figure 11, it can be seen that the ride-sourcing services are used more frequently to travel away from the center than towards the center. Although there are a few zones in the center that have much more arrivals than departures. This might be attributed to different mode choices for trips made in opposite directions. However, this requires further research and additional data sources.

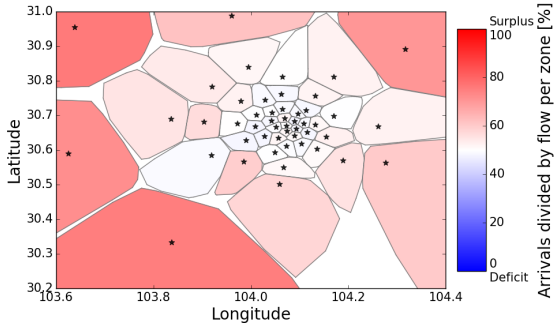


Fig. 11: Share of arrivals in relation to the total flow of the zone.

This can also be observed from the cdf of the relative share of arrivals of all the zones shown in Figure 12. 60% of the zones have more attraction than production. This means that on average the zones with more attraction than production have less rides in comparison to zones with more production than attraction. In other words, zones with more rides have on average more production and zones with less rides have more attraction.

B. Temporal clustering

In this section, the results of the temporal clustering are analyzed using the dendrogram for both different time periods

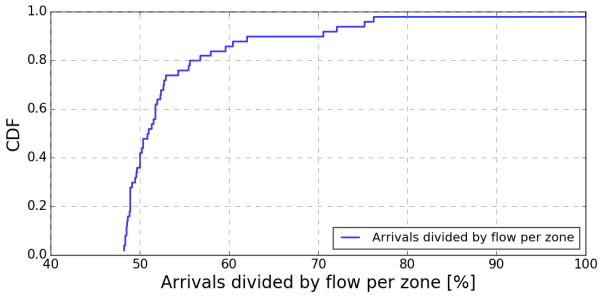


Fig. 12: CDF of the share of arrivals in a zone relative to the total flow in the zone

and different days. Then the different clusters are analysed and labelled according to the demand it represents.

Figure 13 shows the dendrogram of the temporal clustering based on the hourly OD matrices. Since, we considered increments of one hour for the aggregation in the methodology, there are 24 OD matrices and hence 24 leaves in the dendrogram. There are clearly well-defined clusters with similar time periods grouped together. We chose the number of clusters as 5 as it is a comprehensible number of clusters. We can merge or further divide the clusters based on the insights we gain from these clusters.

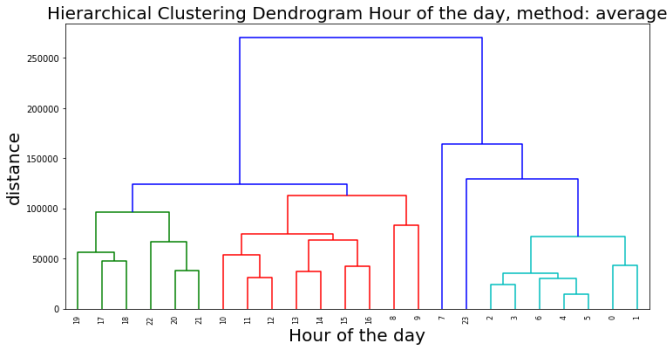


Fig. 13: Dendrogram of the temporal clustering based on the hour of the day

The classes in the dendrogram are numbered from left to right from 1 to 5. The cluster sizes for clusters 1 to 5 are 6,9,1,1 and 7 respectively. There are three large classes and two single cluster classes. Within the large clusters, the distance between the feature vectors differs a lot. This means that within a cluster, some feature vectors have high dissimilarity index whereas others are more similar. A more in-depth look at each cluster and their corresponding medoid is given in Figure 14.

By looking at the distribution of the hours in each cluster in Figure 14, we have labelled each cluster as follows:

- Cluster 1 : **Evening peak hours** starting around 17:00 to 23:00 with high demand and passengers movement from the center towards the neighbouring areas.

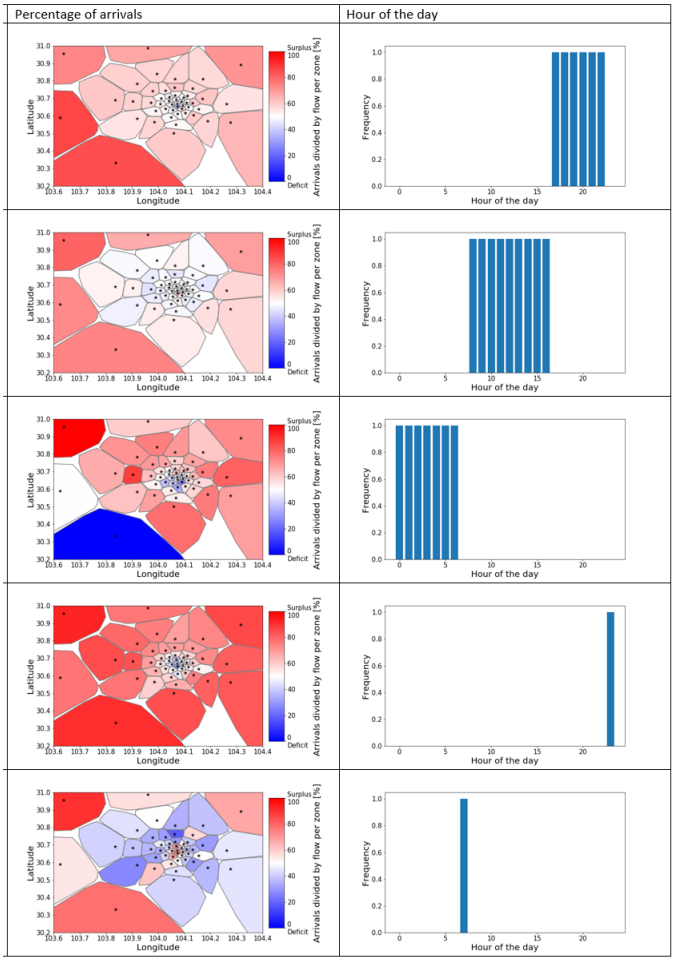


Fig. 14: Medoid and distribution of days in each cluster for temporal clustering based on hourly OD matrix.

- Cluster 2 : **Morning peak hours** starting around 08:00 to 17:00 with high demand and passengers move towards the center and a little towards the outer areas from the neighbouring areas.
- Cluster 3 : **Off peak hours** starting around 00:00 to 07:00 with low demand where passengers move from center towards the outside.
- Cluster 4 : **Transition hour** starting around 23:00 to 00:00 is a distinct hour between the evening peak and the off peak where passengers move from the center towards the outside.
- Cluster 5 : **Transition hour** starting around 07:00 to 08:00 is the distinct hour between the off peak and the morning peak where passengers strongly move towards the center and a little towards the outer areas from the neighbouring areas

There are clear distinct clusters that represents the demand for the different time periods of the day with the demand lowest during the night (cluster 3) followed by the transition periods before and after the night (cluster 4 and 5) and the demand is highest during the day (cluster 1 and 2).

For the temporal clustering for different days, the dendrogram is shown in Figure 15. We chose the number of clusters as 5 for this case as well. The difference between different clusters is comparable to the differences between the different feature vectors within a cluster. There are one large cluster, three small clusters and one single cluster. The large class contains 60% of the days.

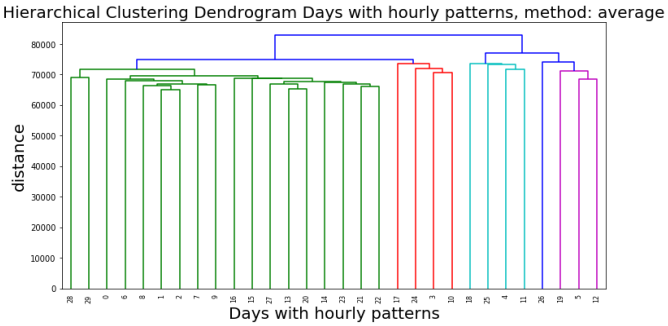


Fig. 15: Dendrogram of the temporal clustering based on the daily patterns

Based the distribution of the days in each cluster and their corresponding medoid shown in Figure 16, the clusters were labelled as follows:

- Cluster 1 : **Monday to Thursday**
- Cluster 2 : **Friday**
- Cluster 3 : **Saturday**
- Cluster 4 : **Sunday**
- Cluster 5 : **Special Sunday**

There is clear division between the demand for different days. There are distinct weekday and weekend patterns. The weekday patterns are further divided with two distinct groups of Monday to Thursday together with demand on Friday different from the normal weekdays. There is also clear distinction between the weekends. All Saturdays in the one month of data are clustered together and the Sundays as well expect one Sunday. Our assumption is that this is a special day given it is being clustered in a group of its own. However, additional information and data is needed to validate this hypothesis. All clusters show similar demand pattern with deficit in the center and surplus in the outer areas with the difference being in the amount of rides for the different days.

V. Conclusion

The results of our clustering analysis suggest that there are pronounced recurrent and thus predictable demand patterns for ride-sourcing services. We find that, differently from demand for private car and public transport, the overall service usage is stable over the day from 09:00 to 21:00 with minor morning and evening peaks and a global peak around 13.30 in the afternoon. The service is mostly used for geodesic distances shorter than 10 km and the typical ride duration is less than 30 minutes. Spatial clustering using the metric inertia resulted in usable zones for further research yielding distinctive types

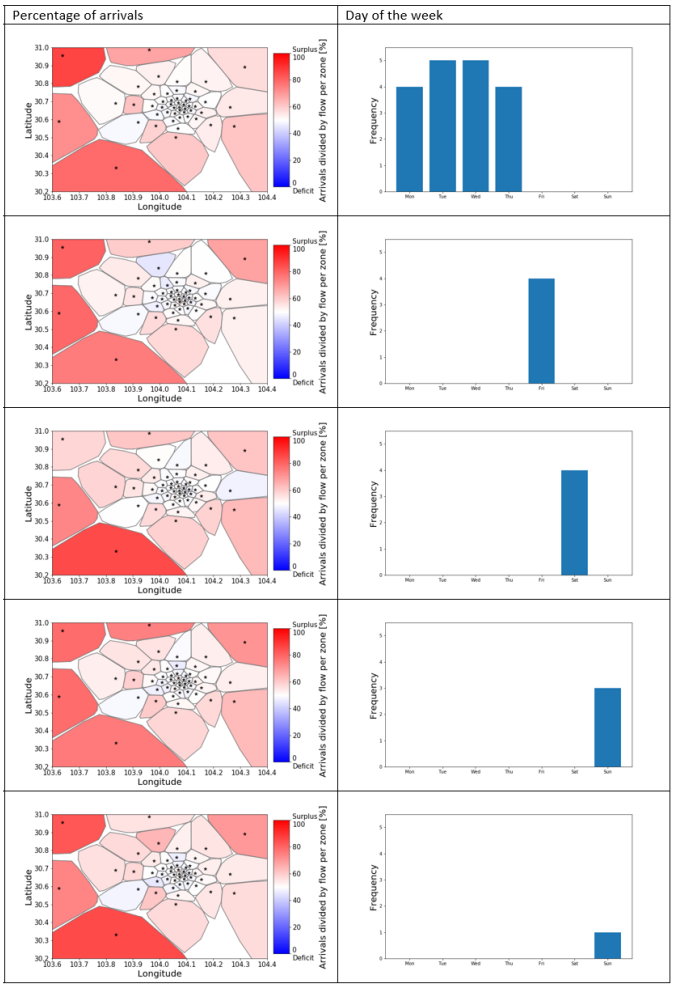


Fig. 16: Medoid and distribution of days in each cluster for temporal clustering based on daily OD matrix.

of zones radiating from the center with increasingly larger geographical area yet lower overall demand level. Ride-sourcing services are mostly used to move either within central zones or from the center to outer neighborhoods (rather than travelling into the center, with the exception of late night hours).

The results of the temporal clustering indicate that the time of the day is most decisive in which class a OD-matrix fits although noticeable differences between weekdays and weekends and recurrent weekly patterns are also manifested. Interestingly, transition hours between the night and morning periods as well as between evening and night periods exhibit a distinctive pattern which differs from the periods preceding and proceeding them. A caveat related to the dataset used is that it contains only orders that have a GPS trace within a pre-defined area, hence under-representing flows within and between outer zones.

While the approach and techniques adopted in this study can be transferred to other contexts and locations, the results of this study are not directly transferable. The dataset pertains to a single month and city. Seasonal variations can thus not be

identified. Another direction for future research is to examine the relation with the service offered by alternative modes, primarily the privately owned car and fixed public transport and their impact on the demand for ride-sourcing services. Also, the clustering could be done taking into account the relative differences between OD-matrices instead of only using the absolute differences. This could give more insights in travel patterns in the city.

The insights gained in this study can aid in developing demand-oriented fleet management for different time periods within a day and between days. Using demand-anticipatory dispatching and rebalancing strategies could improve the service performance as well as the level-of-service by reducing operational costs and shorten waiting times.

ACKNOWLEDGMENT

This research was supported by the CriticalMaaS project (no. 804469) which is financed by a European Research Council and Amsterdam Institute of Advanced Metropolitan Solutions. Data source: DiDi Chuxing GAIA Open Dataset Initiative.

REFERENCES

[1] N. D. Chan and S. A. Shaheen, "Ridesharing in North America: Past, Present, and Future," *Transport Reviews*, vol. 32, no. 1, pp. 93–112, 2012.

[2] Statista, "Statista: Ride-hailing market value worldwide as of May 2018, by key operator (in billion U.S. dollars)," 2018.

[3] Shannon, "Didi now serves 550m users 30m rides per day, growing against Meituan challenges," 2018.

[4] Uber, "Uber Newsroom Company Info," 2018.

[5] E. Yoo, "Didi plans to raise $1.5 billion using asset-backed securities," 2018.

[6] W. Jiang and L. Zhang, "The impact of the transportation network companies on the taxi industry: Evidence from beijings gps taxi trajectory data," *IEEE Access*, vol. 6, pp. 12438–12450, 2018.

[7] S. Guo, Y. Liu, K. Xu, and D. M. Chiu, "Understanding passenger reaction to dynamic prices in ride-on-demand service," in *Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on*, pp. 42–45, IEEE, 2017.

[8] J. Ke, H. Zheng, H. Yang, and X. M. Chen, "Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach," *Transportation Research Part C: Emerging Technologies*, vol. 85, pp. 591–608, 2017.

[9] H. Yao, F. Wu, J. Ke, X. Tang, Y. Jia, S. Lu, P. Gong, J. Ye, and Z. Li, "Deep multi-view spatial-temporal network for taxi demand prediction," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[10] J. Guan, W. Wang, W. Li, and S. Zhou, "A unified framework for predicting kpis of on-demand transport services," *IEEE access*, vol. 6, pp. 32005–32014, 2018.

[11] M. van Engelen, O. Cats, H. Post, and K. Aardal, "Enhancing flexible transport services with demand-anticipatory insertion heuristics," *Transportation Research Part E: Logistics and Transportation Review*, vol. 110, pp. 110–121, 2018.

[12] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA, 1967.

[13] D. Luo, O. Cats, and H. van Lint, "Constructing transit origin–destination matrices with spatial clustering," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2652, pp. 39–49, 2017.

[14] L. Rokach and O. Maimon, "Clustering methods," in *Data mining and knowledge discovery handbook*, pp. 321–352, Springer, 2005.

[15] P. Krishnakumari, A. Perotti, V. Pinto, O. Cats, and H. van Lint, "Understanding network traffic states using transfer learning," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1396–1401, IEEE, 2018.

[16] B. Everitt and A. Skrondal, "The cambridge dictionary of statistics 2002," *Cambridge, Cambridge*.

[17] T. Jun, "The case of Chengdu, China," 2003.

[18] Demographia, "Demographia world urban areas," 2018.

[19] Gaia Didi Chuxing, "https://gaia.didichuxing.com/."