Delft University of Technology

Additional Graduation Work
CIE5050-09

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

Author:
Theo Long Kee Liu (4204387)

Thesis committee:
Dr. O. Cats
P.K. Krishnakumari

January 11, 2019

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

***Abstract*—On-demand transit has become a common mode of transport with ride-sourcing companies like Uber, Lyft, Didi transforming the way we move. With the increase in popularity for such services, the supply needs to adapt according to the demand. For this, the demand needs to analyzed to examine if there are recurrent patterns in them; making it predictable and easily manageable. The identified demand patterns can then be used for optimized fleet management. In this paper, we propose three steps for extracting such demand patterns from travel requests (1) constructing the origin-destination zones by spatial clustering (2) calculating the hourly origin-destination matrix for each day, and (3) temporal clustering to extract the dynamic demand patterns.**

We demonstrate the three step approach on the open-source Didi taxi data. The 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. It can provide insight into the day-to-day regularity and within-day regularity of the demand patterns in Chengdu.

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

I. Introduction

Technical developments of smartphones integrating GPS functionality, internet connectivity and trust in online social networks makes it possible for ride-sourcing to evolve to phase five, a technology-enabled ride-matching [1]. These ride-sourcing companies are creating social networking platforms for real-time ride-sourcing services. The estimated worldwide market value of ride-sourcing services is over 150 billion U.S. dollars in May 2018 with Uber, Didi Chuxing (Didi) and Lyft on top with respectively a valuation of 72, 56 and 11.5 billion U.S. dollars [7]. The first two are considerably larger than the other companies.

In terms of operations, Didi is considerably larger. In 2018 the number of daily rides is 30 million and is twice as high as Uber. Besides, Didi has 550 million users and 21 million drivers, which is about 7 times of Uber with 75 million users and 3 million drivers [6], [8], [9].

A. Problem description

The information about the demand patterns is commercially sensitive, so there is not much literature about it. Didi processes a large amount of data from this ride-sourcing service. In 2017 with about 20 million daily rides, according to Didi over 2,000 terabytes of data was daily processed [5]. The travel requests contains millions of disaggregated origin and destination locations. The dimensionality of the demand computation have to be decreased.

B. Research objectives

On-demand transit has become a common mode of transport with ride-sourcing companies transforming the way we move. With the increase in popularity for such services, the supply needs to adapt according to the demand. For this, the demand needs to analyzed to examine if there are recurrent patterns in them; making it predictable and easily manageable. The identified demand patterns can then be used for optimized fleet management. The patterns are researched for three cases (i) different hours of the month, (ii) different days of the month, and (iii) different hours of the day. These insights can aid in building a demand-oriented fleet management for different time periods within a day and also for different days.

C. Outline

Section II explains the methodology. Section III gives the application. The results for the spatial clustering and for the temporal clustering are given in section IV. Finally, section V concludes with the key findings, limitations, and future research.

II. Methodology

A. Analysis framework

In this paper, we propose three steps for extracting such demand patterns from travel requests (1) constructing the origin-destination zones by spatial clustering (2) calculating the hourly origin-destination matrix for each day, and (3) temporal clustering to extract the dynamic demand patterns. The flowchart (Fig. 1) shows an overview of the inputs and outputs for the spatial and temporal clustering.

B. Spatial clustering

The spatial clustering aims at minimizing inertia (total squared distance from the points to centroids of the zones) while maximizing the flow (number of trips) between the zones. The unique origin and destination points of all the rides are used. The K-means algorithm is used. This is done for a range of number of clusters. For every number of clusters,

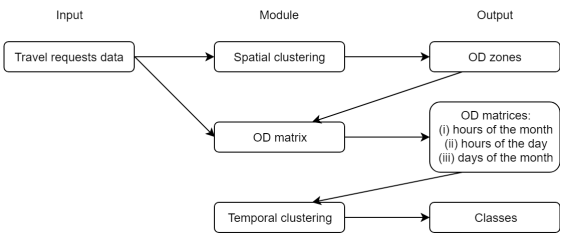

Fig. 1. Flow chart of analysis framework

multiple seeds will be used. The inertia and the intraflow for the different runs are plotted against the number of clusters. The number of clusters is chosen, least as possible, but enough to have a slow decreasing of intraflow percentages for more clusters.

With these static zones as the origins and destinations, we can compute the OD matrix to represent the demand. Each cell in the OD matrix corresponds to the number of trips completed within a time period from a particular origin to a particular destination. The OD matrix can be used to understand where the demand is produced and attracted with respect to the zones and also for extensive descriptive analysis of the demand data.

C. Temporal clustering

For studying this further, we use these OD matrices as feature vectors for the temporal clustering using hierarchical agglomerative clustering. In preparation for the temporal clustering, the OD matrix is split into multiple OD matrices for every individual hour. These are combined into feature vectors for three cases with different levels of aggregation: (i) different hours of the month, (ii) different days of the month, and (iii) different hours of the day. In case ii, the information of the different hours of a day is kept in the feature vector.
The feature vectors of the different cases are clustered together using hierarchical agglomerative clustering.

III. Application

A. Case of Didi in Chengdu

We demonstrate the three step approach on the open-source Didi taxi data. The 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, China, is a city that has been gaining more economical importance over the years as its been rapidly developing and becoming a main hub for several different industries [4].
Like many Chinese cities, Chengdus has a large urban area of more than 1,700 km. 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].
Real travel request data is obtained from Didi Chuxing [3] and is explored in this paper. The ride request data is available for all rides started in November 2016 and have at least one Global Positioning System (GPS) point within the range of

TABLE I
Range of Geographic Coordinates [latitude,longitude]

[30.727818, 104.043333]	[30.726490, 104.129076]
[30.655191, 104.129591]	[30.652828, 104.042102]

TABLE II
Multiple occurrences of orders on all days, including 125 rides with a geodesic distance of higher than 400 km

Occurrences	1	2	3	4	5	6	7	8	9
Frequency	5,144,678	959,782	505	21	10	5	1	0	1

geographic coordinates (Table I) which is a part of Chengdu. The origin and destination points can be outside this area. The data includes origin-destination (OD) points as well as start and end times of the rides. Besides ride request data, route data with GPS data is given (Fig. 2). The orders have unique identification (ID) strings to link the two databases. The route data also includes the driver ID and a time stamp with an accuracy of 2-4 s. In this paper, the route data is not used.

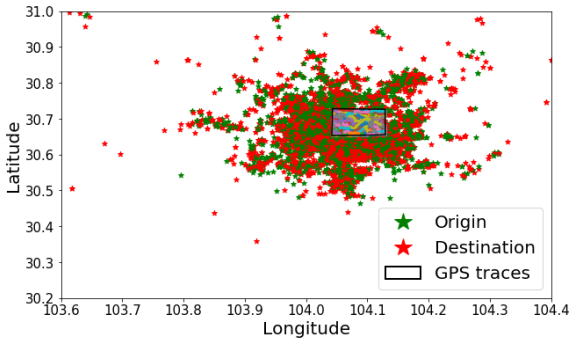

Fig. 2. Spatial plot of the first 25000 rides from the DiDi data with origins, destinations and GPS traces. The real size of the plot is about 76.7x88.7 km (width times height) and the rectangle with the GPS traces is about 8.3x8.1 km (width times height), both widths measured at halfway the latitude.

The dataset is prepared by removing the orders that are repetitions of previous orders with the same order ID. Over 15% of the requests have with multiple occurrences (Table II). The duplicates are removed, since the number of passenger per ride and ride-sharing are not considered. Also, 126 rides with a geodesic distance of higher than 400 km are removed. Rides with very high speeds are kept since it is unclear what the incorrect data is and they provide information on the origin, destination and ride start time. The dataset after preparing the data contains 6,104,877 unique rides.

B. Descriptive statistics

In this section the descriptive statistics of the whole data set are given for the following metrics:

- temporal distribution over the day
- geodesic distance
- travel time
- average geodesic speed, which is the geodesic distance divided by the travel time

The temporal distribution over the day is relatively stable between 09:00-21:00 (3). In comparison to traditional modalities the morning and evening peaks are limited. The morning peak peaks between 09:00-09:30, which is late in comparison to traditional modalities. Here, those traditional rush hour peaks are limited, while the highest peak is between 13:30 to 14:00.

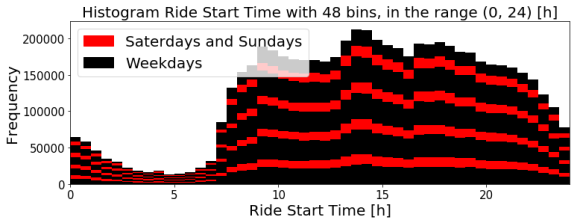

Fig. 3. Histogram Temporal Distribution with 48 bins, in the range (0, 24) [h]

The histogram of geodesic distances shows that the service is mainly used for rides shorter than 10 km (Fig. 4). The peak is between 3-4 km. The average geodesic distance is 6.44 km and the median is 5.28 km. The histogram of travel times shows that most rides are between 5-35 minutes (Fig. 5). The peak is between 10-20 minutes. The average ride duration is 22.12 minutes and the median is 19.28 minutes. The histogram of average geodesic speeds shows most of the rides between 10-30 km/h (Fig. 6). The peak is between 10-20 km/h. The average geodesic speeds is on average 20.03 km/h and the median is 16.41 km/h. These three histograms follow a positive skewed bell shape, meaning the mass is in the shorter rides.

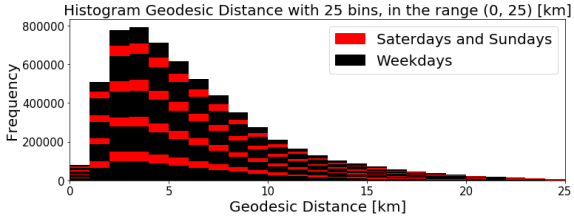

Fig. 4. Histogram Geodesic Distance with 25 bins, in the range (0, 25) [km]

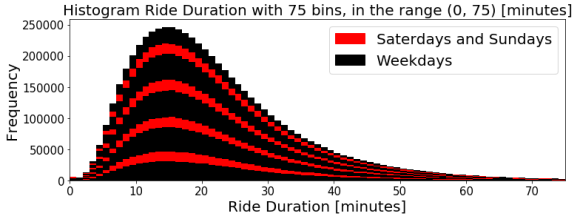

Fig. 5. Histogram Ride Duration with 75 bins, in the range (0, 75) [minutes]

C. Parameters choices

1) Spatial clustering:

- Range of number of clusters: 5 to 100 clusters.
- Number of initial seeds: 10 seeds.
- Maximum number of iterations: 300.

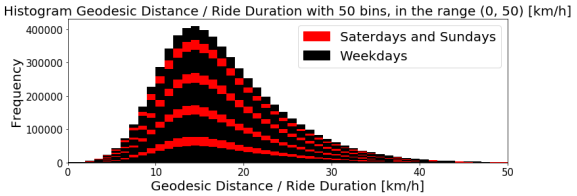

Fig. 6. Histogram Geodesic Distance per Ride Duration with 50 bins, in the range (0, 50) [km per h]

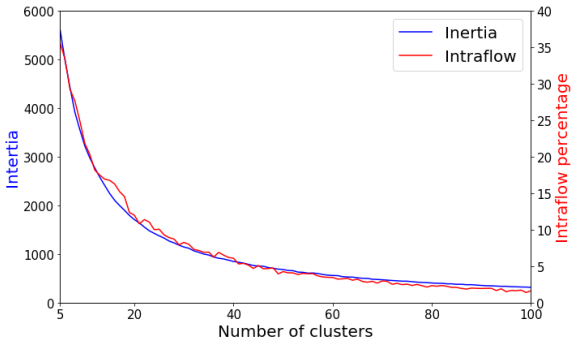

Fig. 7. Inertia and intraflow for 5 to 100 clusters using the runs with respectively the lowest inertia and intraflow. different centroid seeds per number of clusters

2) Temporal clustering:

- Hierarchical clustering method: average.
- Distance metric: cityblock; which uses the sum of the absolute difference between all OD-pairs of different OD-matrices.
- Number of classes: five was chosen to make the analysis informative and comprehensible.

IV. Results

A. Spatial clustering

1) Analysis: The results of the inertia (the summed squared distance between all points and the closest centroid) and intraflow for the different runs are plotted (Fig. 7). Both follow a similar shape, with the inertia and intraflow dropping fast in the first increments of number of clusters and then dropping slower. Since the Kmeans clustering is minimizing the inertia, the line of the inertia is smoother. A number of clusters of 50 is chosen, as a balance between a low intraflow (less than 5%) and a comprehensible number of zones.

2) Resulting zones: This results in 50 zones. A spatial plot with all the rides shows that the zones in the center have up to 500,000 rides per month (Fig. 8). The cumulative density functions (CDF) of the production and attraction (Fig. 9) shows similar functions for production and attraction. The differences between the zones with less rides and zones with more rides is large. The 20% lowest zones have less than 10,000 departures or arrivals in the month November while the top 20% has over 200,000 departures or arrivals. This can also be seen in Fig. 10 and 11 with large color differences between the different zones. The center with small zones

has a much higher number of rides in comparison to the neighbouring zones.

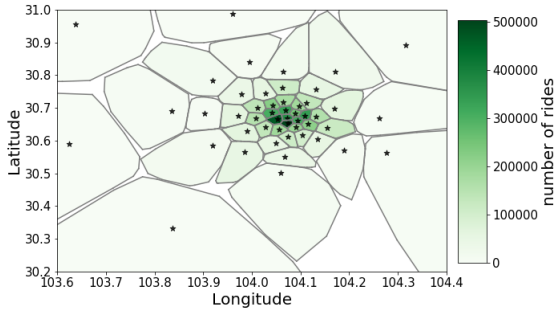

Fig. 8. Spatial plot of the number of rides per zone

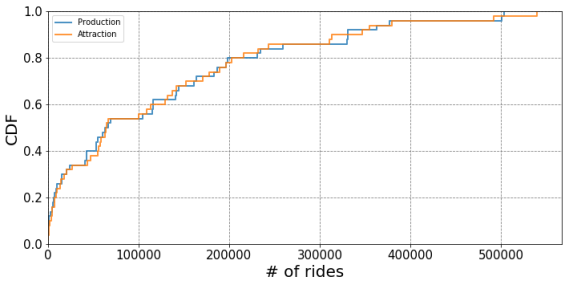

Fig. 9. CDF production and attraction

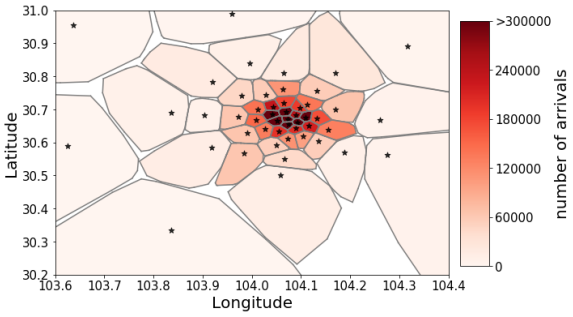

Fig. 10. Spatial plot of the number of arrivals per zone

Based on this, the zones are divided into five rings:

- Ring 1: the center with small zones and very high demand
- Ring 2: the neighbouring zones with high demand
- Ring 3 and 4: the medium zones with medium demand
- Ring 5: the outer areas with large zones with low demand

The intraflow calculated per zone as the OD-pair from and to that same zone and divided by the total production of that zone. In over 60% of the zones the intraflow is less than 1%. In about 30% of the zones, the intraflow is higher than 4%, but always lower than 10% (Fig. 13). The intraflow is highest in the zones where the GPS traces are (Fig. 14). This could

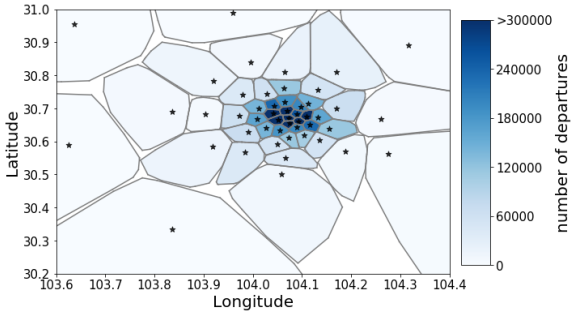

Fig. 11. Spatial plot of the number of departures per zone

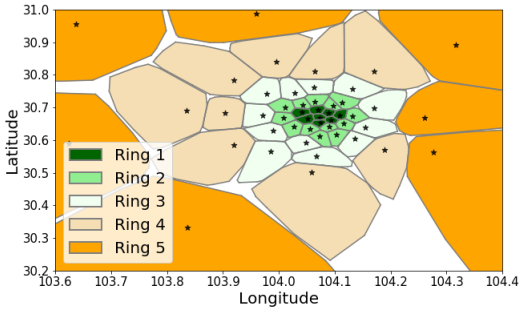

Fig. 12. Spatial plot of the different rings

mean that ride-sourcing services are also used between 4 and 10% of the time for short rides within zone rides.

60% of the zones have more attraction than production (Fig. 15 and 17). 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. This phenomenon is observed in Figures 16 and 18. This means that ride-sourcing services are more used 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.

B. Temporal clustering

In this section the results are analyzed with the use of the dendograms. Secondly, the resulting classes are characterized.

1) Dendogram analysis: The classes in the dendograms are numbered from left to right from 1 to 5. The metric distance is the sum of the absolute difference between the OD-pairs of the different OD-matrices. If clusters of OD-matrices are compared, the average OD-matrix is used per cluster. The cluster sizes for different classes and cases is given in Table III.

Case (i) hour of the month has three large classes and two small classes (Fig. 19). Class 1 has the lowest distances, caused largely by the low number of orders during those hours. The distance to the other classes is large, meaning class 1 is distinctive.

Case (ii) day of the month has one large class, three small

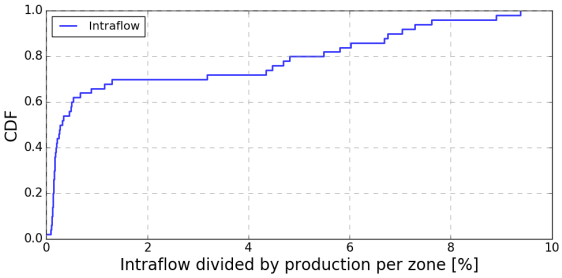

Fig. 13. CDF intraflow

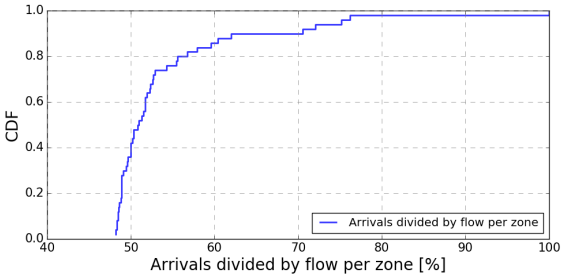

Fig. 15. CDF share of arrivals in comparison to the flow in a zone

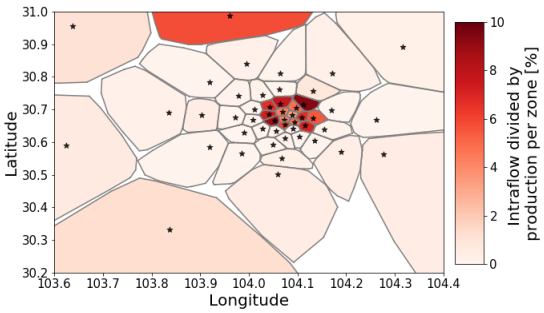

Fig. 14. Spatial plot intraflow

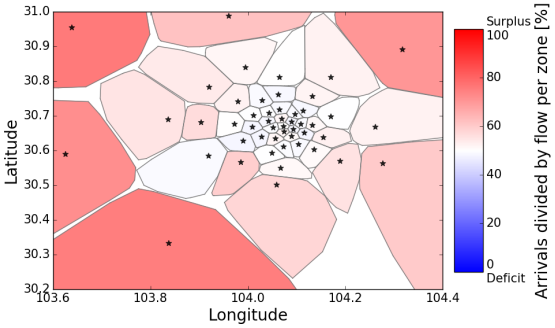

Fig. 16. Spatial plot share of arrivals

classes and one single cluster class (Fig. 20). The large class consists of 60 percent of the days. The differences between different clusters is comparable to the differences between the different feature vectors within a cluster.

Case (iii) hour of the day has three large classes and two single cluster classes (Fig. 21). Within the large classes, the distances between feature vectors differs a lot. This means that within a class, some vectors are very much alike and others are much less alike.

2) Resulting classes: This section presents the results of the clustering for the three cases. The characteristics of the classes per case are given. The plots and interpretation of the number of arrivals and departures and the percentage of arrivals are based on a medoid per class.

For case (i), four levels are used to describe the number of arrivals and departures: low, medium, high, very high. These roughly corresponds with the following amount of arrivals or departures per zone per hour: less than 300, between 300 and 700, between 700 and 1000, and over 1000. The five classes for case (i) hour of the month (visualized in Fig. 22):

1) The demand is low. The general movement is slightly away from the center. This class characterizes the patterns for every day of the week between 00:00-07:00 and also for every day of the week except Friday and Saturday between 23:00-00:00.
2) The number of arrivals is high to very high in the center. The number of departures is high to very high in the center. The general movement is from the neighbouring zones towards the city center and the outer areas. This class characterizes the patterns for every day of the week between 11:00-17:00 and also for the weekdays between 10:00-11:00 .
3) The number of arrivals is medium in both the center and the neighbouring zones. The number of departures is high to very high in the center. The general movement is from the center to the neighbouring zones and the outer areas. This class characterizes the patterns for every day of the week between 17:00-23:00 and also for Friday and Saturday between 23:00-00:00.
4) The number of arrivals is very high in the center. Compared to class 3, the arrivals in the neighbouring zones are lower. The number of departures is high to very high in the center zones. This class characterizes the patterns for every weekday between 08:00-10:00.
5) The number of arrivals are high in the center zones. The number of departures is medium in the center zones and the neighbouring zones. The general movement is from the neighbouring zones and the outer areas towards the center. This class characterizes the patterns for every day of the week between 07:00-08:00 and also for Saturday and Sunday between 08:00-10:00.

For case (ii) different days of the month (Fig. 23), the key

TABLE III
CLUSTERSIZES OF THE DIFFEREN CLASSES FOR THE THREE CASES

Case (i)	232	202	188	44	54
Case (ii)	18	4	4	3	1
Case(iii)	6	9	7	1	1

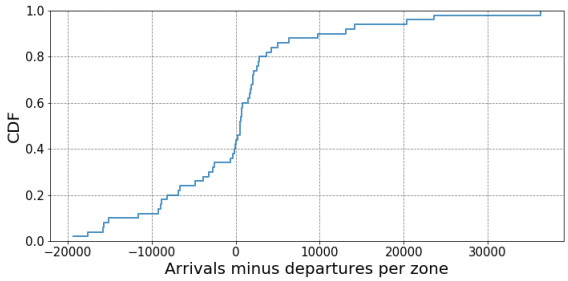

Fig. 17. CDF of the number of arrivals minus departures; positive means surplus, negative means deficit

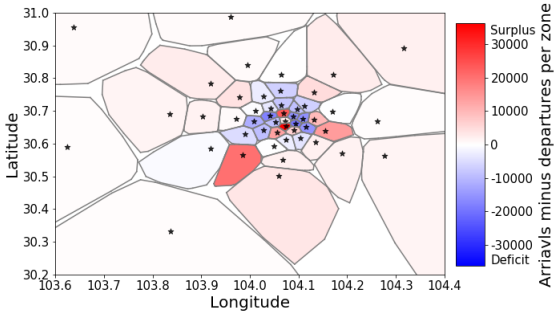

Fig. 18. Spatial plot of the number of arrivals minus departures

findings are:

- There is a clear division between Monday to Thursday, which is the largest class, and the other days. Monday to Thursday are more similar than the days in the weekend.
- All days show the same pattern with a deficit in the center and a surplus in the outer areas.

For case (iii) different hours of the day (Fig. 24), the key findings are:

- Since every hour of the day is viewed as one feature vector, every hour of the day will only be in one class.
- The different classes are distinctive in terms of demand. The demand is lowest during the night (class 3). Followed by the transition periods before and after the night (class 4 and 5). The demand is highest during the day (class 1 and 2).
- The different classes have different movement patterns. Between 23:00-00:00 passengers go from the center towards the outside (class 4). This pattern continues, but less strong, during the night until 07:00 (class 3). At the beginning of the morning peak, passengers strongly move towards the center and a little towards the outer areas from the neighbouring areas (class 5). During the day this decreases in strength, but continues (class 2). The movement from the center towards other areas starts around 17:00 (class 1).

V. Conclusion

The key findings of this paper are:

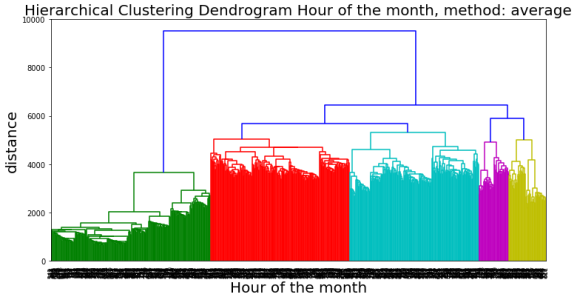

Fig. 19. Dendogram hour of the month, method: average

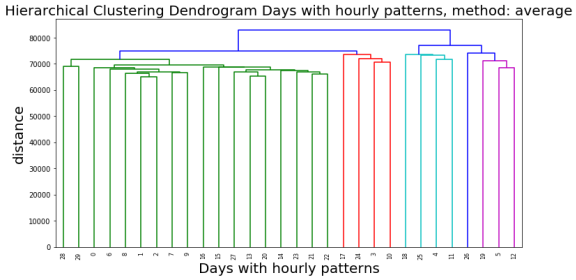

Fig. 20. Dendogram day of the month with hourly patterns, method: average

- The usage of this service is stable over the day from 09:00 to 21:00. Morning and evening peaks are small. The highest usage around 13:30.
- The service is mostly used for geodesic distances shorter than 10 km and the ride duration is less than 30 minutes.
- Spatial clustering using the metric inertia resulted in usable zones for further research. There are different distinctive types of zones; center zones which are small and heavy used, neighbouring zones which are a bit larger and medium used and outer areas which are large zones and a lightly used.
- Zones with more rides have on average more production and zones with less rides have more attraction. This could mean that ride-sourcing services are more used to move away from the center than to move towards the center.
- The temporal clustering for case (i) different hours of the month showed that the time of the day is mostly deciding in which class a OD-matrix fits. Although there are some differences between weekdays and weekends, and also for the hour 23:00-00:00 for Friday and Saturday.
- In case (i), all the hours of the same day of the week and same hours of the day fall into the same class, which means that the demand pattern is weekly repetitive.
- In case (ii) is a clear division between Monday to Thursday, which is the largest class, and the other days. Monday to Thursday are more similar than the days in the weekend.
- In case (iii), the different classes are distinctive in terms of demand. The demand is lowest during the night. Followed by the transition periods before and after the night. The demand is highest during the day.

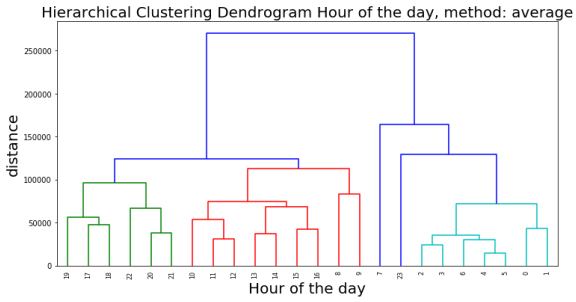

Fig. 21. Dendogram hour of the day, method: average

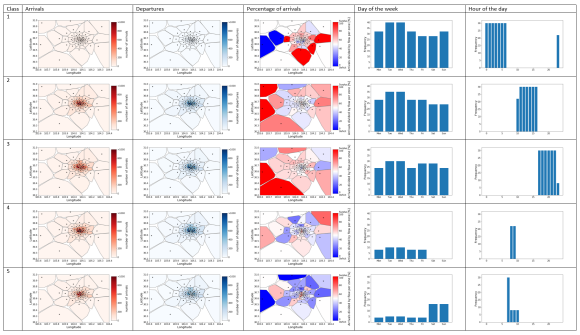

Fig. 22. Table case (i) hour of the month with Different Classes and characteristics

- In case (iii), the different classes have different movement patterns. Between 23:00-00:00 passengers go from the center towards the outside. This pattern continues, but less strong, during the night until 07:00. At the beginning of the morning peak, passengers strongly move towards the center and a little towards the outer areas from the neighbouring areas. During the day this decreases in strength, but continues. The movement from the center towards other areas starts around 17:00.
- These insights can aid in building a demand-oriented fleet management for different time periods within a day and also for different days.

There are some limitations in this research. The data acquired is of one month in one city. Even though this dataset cannot provide insight into the seasonal demand pattern variations, it can provide insight into the day-to-day regularity and within-day regularity of the demand patterns in Chengdu. It is unclear how the actual performance of other modalities influence the usage of this ride-sourcing service. Results could vary depending on the choices made for the parameters. However, the results seem plausible. A bias in the data is that it contains only orders that have a GPS trace within a limited area. This could bias the results for intraflow. However, this paperś main goal is to characterize and cluster demand patterns throughout time, which is still possible.

Based on this research the fleet size could be managed based on the time and day. Using rebalancing strategies could improve the companies performances. In future research, short-term prediction could be explored. Also, the clustering could

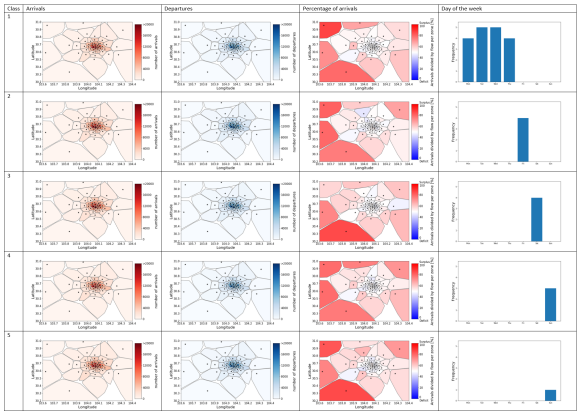

Fig. 23. Table case (ii) days with hourly patterns with Different Classes and characteristics

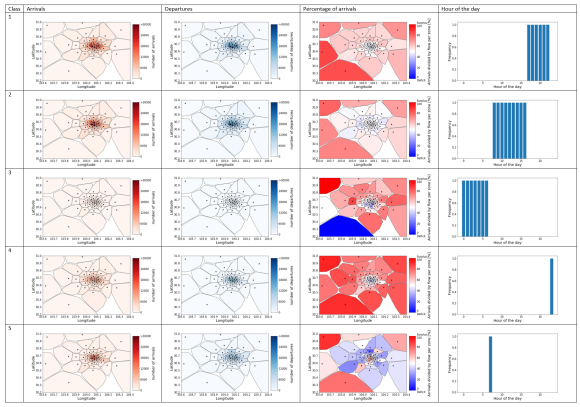

Fig. 24. Table case (iii) hour of the day with Different Classes and characteristics

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. Also research could be done on more data, to see how these results fits in other places or with seasonality influences.

REFERENCES

[1] Nelson D. Chan and Susan A. Shaheen. Ridesharing in North America: Past, Present, and Future. *Transport Reviews*, 32(1):93–112, 2012.
[2] Demographia. Demographia world urban areas, 2018.
[3] Gaia_Didi_Chuxing. Gaia Didi Chuxing.
[4] Tian Jun. The case of Chengdu, China, 2003.
[5] Paul Sawers. How Chinas meshing ride-sharing data with smart traffic lights to ease road congestion, 2017.
[6] Shannon. Didi now serves 550m users 30m rides per day, growing against Meituan challenges, 2018.
[7] Statista. Statista: Ride-hailing market value worldwide as of May 2018, by key operator (in billion U.S. dollars), 2018.
[8] Uber. Uber Newsroom Company Info, 2018.
[9] Eva Yoo. Didi plans to raise $1.5 billion using asset-backed securities, 2018.