Incorporating Multi-Source Urban Data for Personalized and Context-Aware Multi-Modal Transportation Recommendation

APPENDIX A
EXTENDED DATA DESCRIPTION

We first introduce three kinds of user behavior data in detail, include *query record*, *display record* and *click record*.

- **Query record.** A query record consists of a session ID, an anonymized user ID, a time stamp, the coordinates and the POIs of the origin o and the destination d, and the operating system of the device. For example, $[s_i, u_i$, "$2018-09-01\ 15:15:36$", $(116.30, 40.05)$, $(116.353, 39.99)$, "$Baidu\ Building$", "$Beihang\ University$", "IOS"$]$ means a user u_i makes a query on a trip from Baidu Building to Beihang University in the afternoon of September 1st, 2018.
- **Display record.** A display record consists of a session ID, an anonymized user ID, a time stamp and a list of routes. Each route consists of the transport mode, the estimated route distance, the estimated time of arrival (ETA), the estimated price and the rank in the display list. The number of displayed routes varies across queries, and there can be no feasible routes for certain queries.
- **Click record.** A click record consists of a session ID, an anonymized user ID, a time stamp, and a list of clicked routes in the route list. There can be none or multiple clicks on a route. We only record the first click on each route and remove repeated clicks.

Then we present three kinds of geographical data in detail, include *POI data*, *road network data* and *transportation station data*.

- **POI data.** Semantics in POIs indicate the travel intention and have been applied for various urban computing tasks. Our POI dataset contains $1,204,344$ distinct POIs in BEIJING and $1,594,684$ distinct POIs in SHANGHAI. Each POI record has a POI ID, an ascendant POI ID, coordinates of the location, the POI name and a two-level category. The ascendant POI is the higher level POI of the current POI. For example, "Baidu building" is the ascendant of "Baidu building tower 2". To enrich the POI semantics, we map uninformative POI categories such as "Entrance" and "Door Address" to the ascendant POI categories. The two-level category has a primary category and a secondary category. For example, "Education" is a primary category whereas "University" is one of its secondary categories. There are 18 primary categories and 189 secondary categories in the POI dataset.
- **Road network data.** Road network data help to capture regional traffic capability. Each record of road network consists of a unique road segment ID, the start location coordinates, the end location coordinates, the road length and the level of the road segment. There are eight levels of road segments. For instance, the national highway is with the highest level and the pedestrian path is with the lowest level.
- **Transportation station data.** The distribution of transportation stations also influences user preferences on transportation modes. For regions with few bus stations, taxis might be preferred. Each record of transportation station data consists of a unique station ID, coordinates of the station location, a list of bus lines across the station and the corresponding city code.

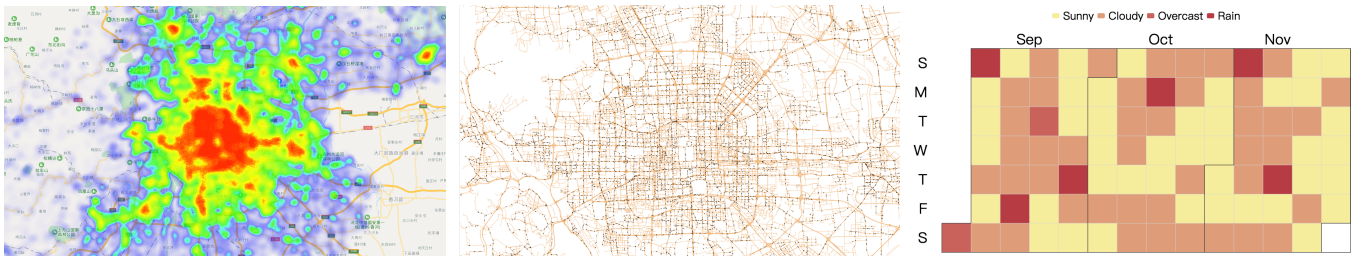
TABLE 1
Statistics of Primary POI Categories of BEIJING.

ID	Category	Count	ID	category	Count
P01	Residence	163,733	P10	Healthcare	16,123
P02	Shopping	137,882	P11	Finance	14,688
P03	Company	137,223	P12	Entertainment	13,894
P04	Entrance	110,667	P13	Hotel	12,700
P05	Life Service	85,448	P14	Culture Venue	9429
P06	Food	78,088	P15	Sports	9,022
P07	Government	37,546	P16	Tourist Attraction	7,763
P08	Education	35,035	P17	Door Address	7,709
P09	Beauty	19,650	P18	Administrative area	4,069

Table 1 shows the distribution of primary POI categories. The spatial distribution of POIs in BEIJING is show in Figure 1(a). Figure 1(b) shows the spatial distribution of road networks and transportation stations, where the yellow lines are road segments and black points are stations. Similar to user activities, the density of POIs, road segments and bus stations in the urban central area is much higher. Figure 1(c) shows the distribution of weather in each day. Overall, there are more rainy days in September and November whereas more sunny days in October.

APPENDIX B
DETAILED FEATURE LIST

Table 2 is the feature list used in Hydra, including *Plan features*, *Spatial features*, *Temporal features*, *Meteorological features*, and



(a) POI distribution. (b) Road network and station distribution. (c) Weather distribution.

Fig. 1. More distributions of the BEIJING dataset: (a) the distribution of POIs; (b) the distribution of road networks and stations; (c) the distribution of weather

TABLE 2
The Description of features.

Feature Type	Feature	Description
Plan	Road network distance	The length of the planned route on the road network
	Price	The total cost of the plan
	ETA	The estimated time of arrival (ETA) of the plan
	Transfer count	The number of transfers of the plan
	Transport mode count	The number of transport modes used in the plan
Spatial	District	The administrative district which the origin and destination belongs to
	POI category	The primary and secondary category of the POI
	Spherical distance	The spherical distance between the OD pair
	Station distance	Spherical distances of top-k nearest bus stations from the O/D location
	Station count	The number of bus stations in the O/D region
	Regional POI distribution	The distribution of two level POI category of corresponding O/D region
	Regional road network distribution	The number of road segment and road intersection in the O/D region
	Regional bus line count	The number of bus line cross the O/D region
	Regional historical mode distribution	The mode preference distribution of O/D/OD region
Temporal	Hour	The corresponding time period in a day
	Minute	The corresponding minute bin
	Day of week	The ordinal number of the day in a week
	Day of month	The ordinal number of the day in a month
	Workday	Whether the day is a workday
Meteorological	Weather	The weather in current time period
	Temperature	The temperature and statistics (i.e., highest/lowest temperature) in current day
	AQI	The AQI and AQI statistics (i.e., highest/lowest AQI) in current day
	Wind speed	The wind speed in current time period
	Wind direction	The wind direction in current time period
User	Demographic attribute	The age, gender of the user and OS in use
	Social attribute	The education level, industry type, car type and consumption level
	User historical mode	The mode preference distribution of the user

User features.

APPENDIX C
SPHERICAL DISTANCE CALCULATION

Given (φ_1, λ_1) as the coordinates of origin o and (φ_2, λ_2) as the coordinates of destination d. The spherical distance of od is calculated as follows:

$$d_{od} = 2R \cdot atan2(\sqrt{\sin^2(\Delta\varphi_{od}/2) + \cos\varphi_1 \cos\varphi_2 \sin^2(\Delta\lambda_{od}/2)}, \sqrt{1 - \sin^2(\Delta\varphi_{od}/2) - \cos\varphi_1 \cos\varphi_2 \sin^2(\Delta\varphi_{od}/2)}) \quad (1)$$

Where $\Delta\varphi_{od} = \varphi_1 - \varphi_2$ and $\Delta\lambda_{od} = \lambda_1 - \lambda_2$. We set $R = 6371$ to approximate the distance of od on the earth's surface.

APPENDIX D
DATA INTEGRATION

We integrate multi-source urban datasets into a unified dataset to create a more comprehensive view of transportation mode choices. Specifically, by integrating the user behavior data and the meteorological data, we can find the key weather factor influencing users' transport mode preference. As another example, by integrating the user behavior data and POI data, we can analyze the relationship between transport modes and travel intention.

We use the $JOIN$ operator to integrate all datasets together. We first join user queries, display and click records on the session ID. Since each query contains an origin and a destination, we further join the origin and the destination with the POI data through location coordinates and POI names. Note that in map search queries, 91% origins are "current locations", which do not have explicit POI names. For such origins, we associate the coordinates with the nearest POI. Besides, the meteorological data is in district level. Therefore, we join the origin and the destination with the meteorological record if the coordinates located in the corresponding district. We crawl the polygon of each district as a preprocessing step and then join the original and destination with the meteorological record if the coordinates located in the corresponding district. Finally, we join the above data with user profile data through user IDs.