Supplementary Material for

Measuring Economic Activity in China with Mobile Big Data

Lei Dong*[1,2,3], Sicong Chen[2], Yunsheng Cheng[2], Zhengwei Wu[2], Chao Li[2], and Haishan Wu†[2]

[1]Institute of Remote Sensing and Geographical Information Systems, Peking University, Beijing, 100871, China
[2]Big Data Lab, Baidu Research, Baidu, Beijing, 100085, China
[3]School of Architecture, Tsinghua University, Beijing, 100084, China

Data Access

All indices mentioned in the main text could be accessed via http://bdl.baidu.com/mobimetrics, or the Bloomberg Terminal.

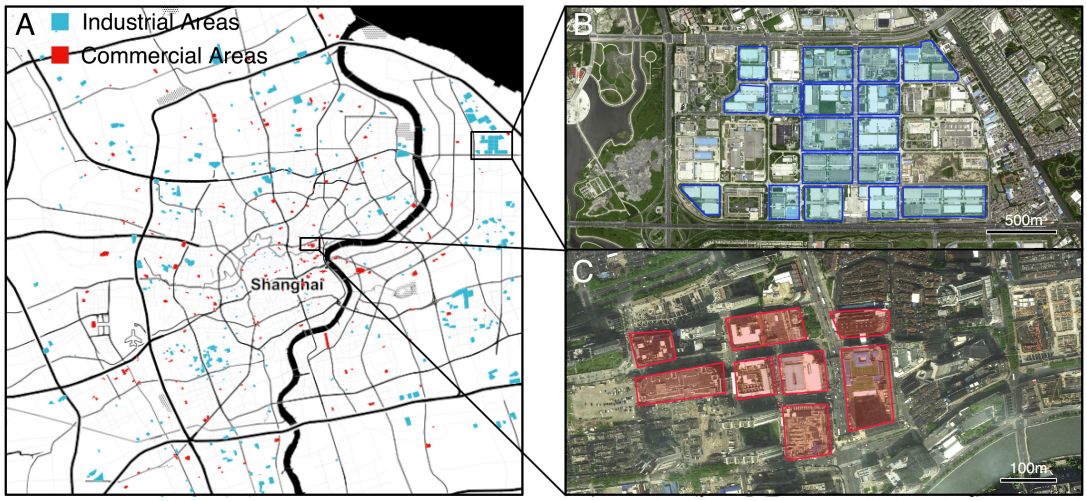
Tags of POI

POI, or Points of Interest, refer to the geo-position points in maps. POI provided by Baidu Maps come from different sources. To assure the quality of POI, Baidu checked the validity of every POI and merged them into a standardized dataset. By comparing names and spatial coordinates of POI, calculating similarities and hashing, Baidu built its own POI dataset with a precision higher than 95%. All the POI were assigned with correct tags that described the attributes of locations. Each POI belongs to a standardized tag, and there are 30 categories and about 240 subcategories in total. In our research, we chose categories related to consumer activities: such as *Shopping*, *Financial Institutions*, *Hotels*, *Tourism*, *Restaurants*, *Auto Services*, and *Entertainments*.

Spatial Distribution of AOI

Areas of Interest (AOI), corresponding to a specific region. In order to measure the employment and consumer activities, we manually labelled about 6,000 AOI, including 2,000 industrial parks and 4,000 commercial areas. Supplementary Figure 1 shows the spatial distribution of AOI of Shanghai. Industrial areas are coloured in blue and commercial areas are in red.

*arch.dongl@gmail.com
†hswu85@gmail.com

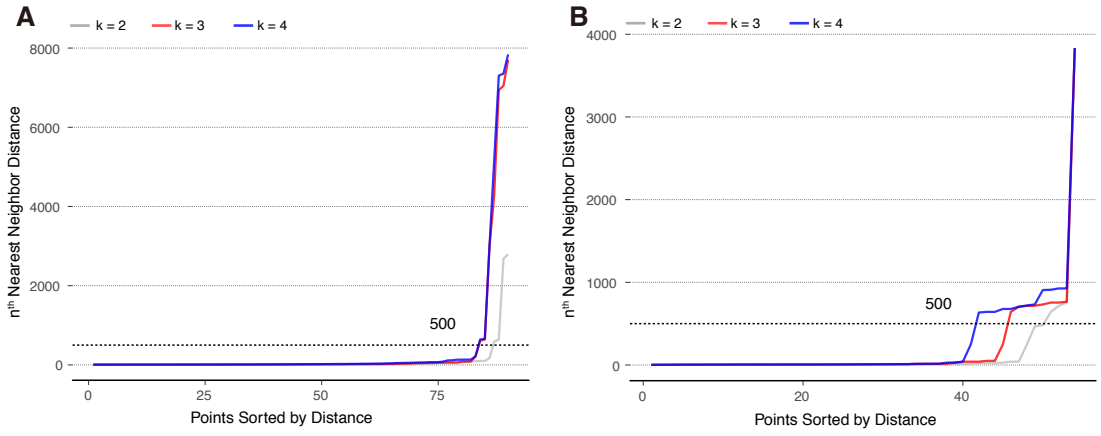


Supplementary Figure 1: (A) Illustration of the spatial distribution of labelled AOI. Industrial areas are coloured in blue and commercial areas are in red. Zoomed-in part: (B) and (C) are industrial parks and commercial areas, respectively. Maps were created by QGIS 2.12 (http://qgis.org/).

Parameters of DBSCAN

Eps and *MinPts* are two important parameters for DBSCAN algorithm. Here we use a data-driven method proposed in (Tan, 2006) to estimate the best parameters. We compute the distance from a point to its k^{th} nearest neighbor (note as $k-dist$), and sort all data points by their $k-dist$ in increasing order. A sharp change may occur at the value of $k-dist$ that corresponds to an appropriate threshold of *Eps*.

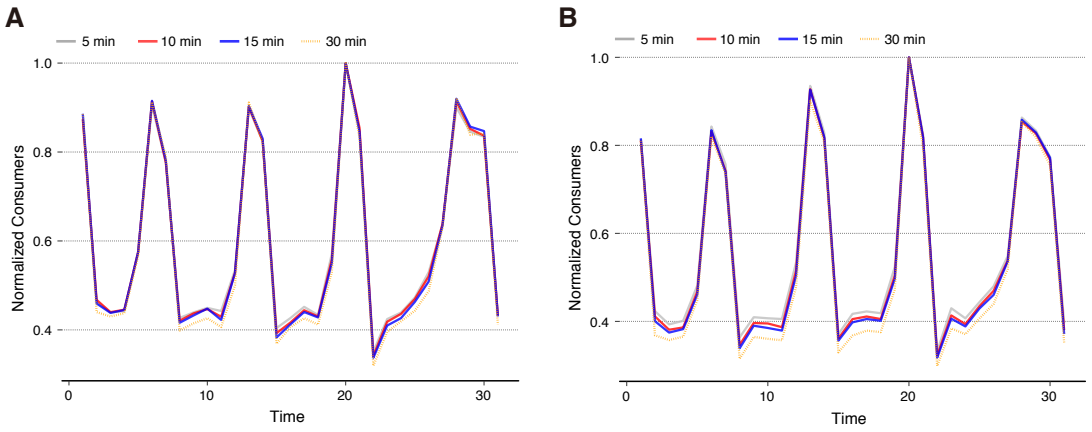
Human mobility traces depict similar patterns, and Supplementary Figure 2 shows two users' analysis results. We could see that $500m$ is a suitable value of as the *Eps* parameter (marked by dotted line). This parameter does not change a lot as k changes.



Supplementary Figure 2: Parameters of DBSCAN. (A) User I. (B) User II.

Parameters of Consumer Detection

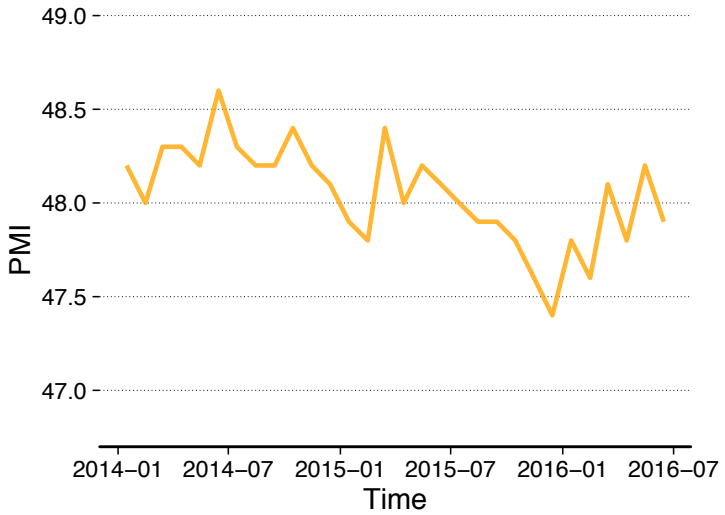
We use 10 minutes as the threshold to define a visit to a place in the main context. To check whether the analysis change by varying this threshold, we set 5, 10, 15, 30 minutes as the threshold respectively. Supplementary Figure 3 shows that the selection of threshold has little impact to the results.



Supplementary Figure 3: Time threshold and consumer detection. (A) AOI I. (B) AOI II.

Manufacturing Purchasing Management Index (PMI)

There are two main PMI data sources in China: One is published by National Bureau of Statistics (NBS), the other one is called by Caixin PMI. In Supplementary Figure 4, we show employment index derived from the NBS manufacturing PMI.



Supplementary Figure 4: National Bureau of Statistics Manufacturing Purchasing Management Index (Employment Index) [Data source: http://data.stats.gov.cn].

The List of Selected Apple Stores

We selected POI of Apple Stores (China) based on the store list from Apple's official website: https://www.apple.com/cn/retail/storelist/. Since Apple kept on expanding its offline stores in China, we only selected those opened before 2015y. The final list was as follows:

- Beijing (Sanlitun, Dayuecheng, Wangfujing, Huamao)
- Shanghai (Guojin Center, Nanjing East Road, HongKong Plaza)
- Tianjin (Henglong Plaza)
- Chengdu (Wanxiangcheng)
- Wuxi (Henglong Plaza)
- Chongqing (Beichengtianjie)
- Shengzhen (Yitian Holiday Plaza)

The quarterly financial results were downloaded from: http://investor.apple.com/.

Map Query and Box Office

The baseline model for box office prediction is auto regression with 1 day and 7 day lags:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-7} + e_t \tag{1}$$

where y_t, y_{t-1}, and y_{t-7} are the revenues of box office at time t, $t-1$, and $t-7$, respectively. And the map query model is the baseline model plus map query variables:

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-7} + \beta_3 q_t + \beta_4 q_{t-1} + \beta_5 q_{t-7} + e_t \tag{2}$$

where q is the normalized (by page view) volume of map queries for cinemas.

Results are shown in Supplementary Table 1. Map query improves the R^2 of the baseline model from 0.489 to 0.934.

Map Query Trends

As stated in the main text, the proposed map query trends are valuable for not only improving forecasting accuracy, but also strengthening policy suggestions, which could be further investigated through cross validation with more data sources.

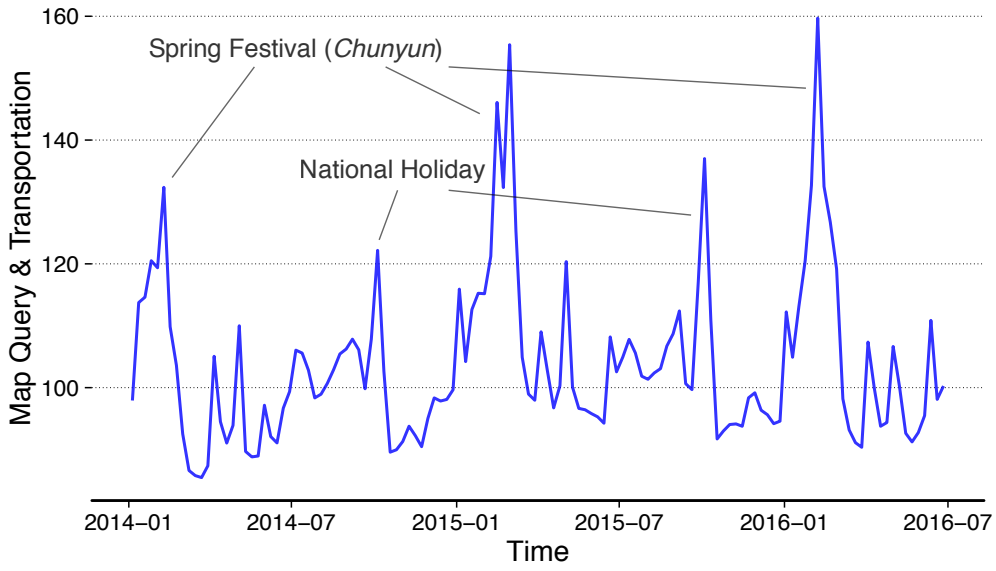
We plot weekly map query trends of Transportation, Hospital, and Fitness.

Variables	Baseline Model	+ Map Query
y_{t-1}	0.557***	0.805***
	(0.041)	(0.035)
y_{t-7}	0.264***	0.083
	(0.041)	(0.034)
q		0.471***
		(0.010)
q_{t-1}		-0.348***
		(0.019)
q_{t-7}		-0.045***
		(0.017)
Observations	353	353
R-squared	0.489	0.934

Note: OLS regressions are estimated with a constant that is not reported in this table. Standard errors are shown in brackets. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

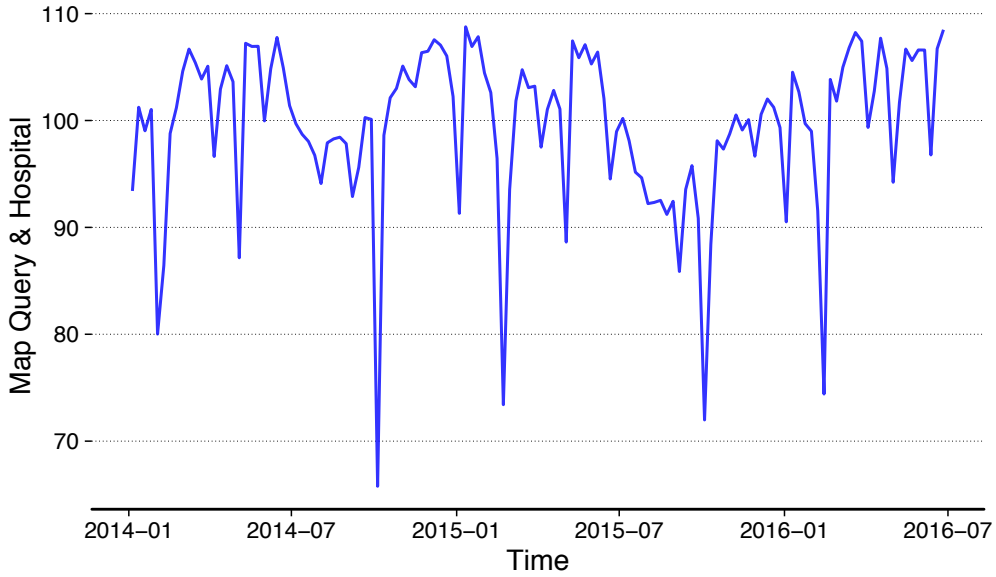
Supplementary Table 1: Regressions of foot traffic and map query

-Transportation



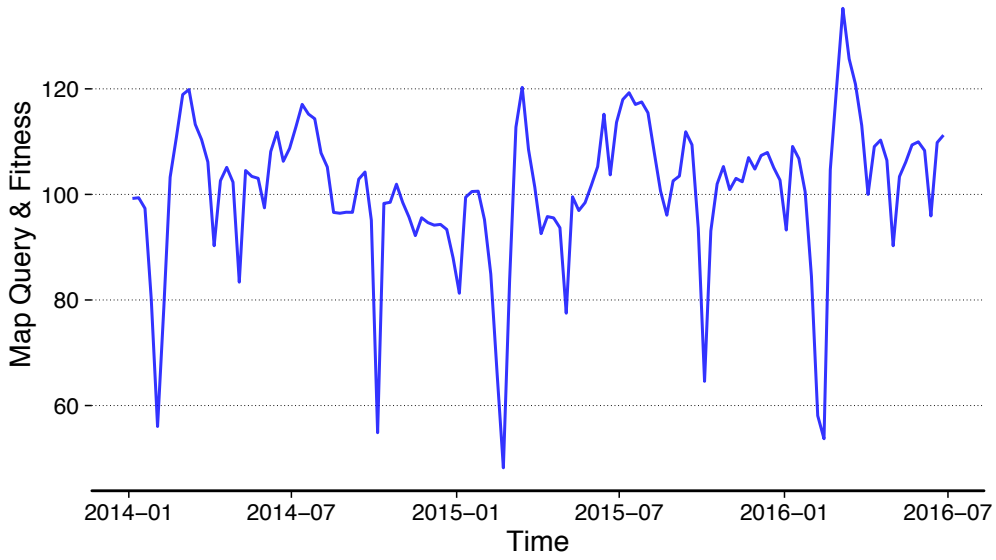
Supplementary Figure 5: Transportation Map Query Trends.

-Hospital



Supplementary Figure 6: Hospital Map Query Trends

-Fitness

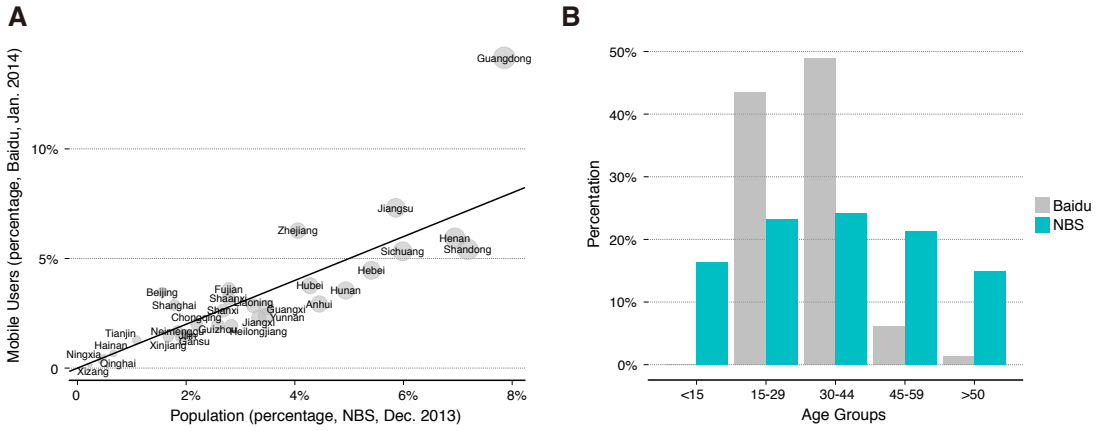


Supplementary Figure 7: Fitness Map Query Trends

Mobile Phone Data versus Official Statistics Demographics

Here, we show the demographic differences between mobile phone data and official statistics. We calculated Baidu users' spatial distribution of January, 2014 based on a sample size of 157 millions we also computed the users' age structure based on a sample of 0.5 million registered users. The corresponding official statistics were derived from the website of National Bureau of Statistics of the People's Republic of China (NBS, http://www.stats.gov.cn/tjsj/).

Supplementary Figure 8 indicates that the spatial distribution of Baidu users is relatively consistent with that of the Chinese population. While, the penetration rate of Baidu users is higher in developed regions, e.g. Beijing, Shanghai, Guangdong, Jiangsu, Zhejiang, as compare to underdeveloped regions, e.g. Anhui, Hunan, Gansu, etc. For age structure, the group of 15-45 accounts for over 90% of the total users in our data sample, and that age group accounts for less than 50% of the China population. In other words, the data over-represent those aged 15-45, and under-represent people in less than 15 age group and over 45 groups.



Supplementary Figure 8: Mobile phone data versus official statistics demographics. (A) Spatial distribution at province level. (B) Age structure.

References

Tan PN (2006) Introduction to Data Mining. Pearson Education, India.