Dong et al. *EPJ Data Science* (2017) 6:29
DOI 10.1140/epjds/s13688-017-0125-5

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

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*]

*Correspondence:
hswu85@gmail.com
[2]Big Data Lab, Baidu Research, Baidu, Beijing, 100085, China
Full list of author information is available at the end of the article

Abstract

Emerging trends in the use of smartphones, online mapping applications, and social media, in addition to the geo-located data they generate, provide opportunities to trace users' socio-economic activities in an unprecedentedly granular and direct fashion and have triggered a revolution in empirical research. These vast mobile data offer new perspectives and approaches to measure economic dynamics, and they are broadening the social science and economics fields. In this paper, we explore the potential for using mobile data to measure economic activity in China from a bottom-up view. First, we build indices for gauging employment and consumer trends based on billions of geo-positioning data. Second, we advance the estimation of offline store foot traffic via location search data derived from Baidu Maps, which is then applied to predict Apple's revenues in China and to accurately detect box-office fraud. Third, we construct consumption indicators to track trends in various service sector industries and verify them with several existing indicators. To the best of our knowledge, this is the first study to measure the world's second-largest economy by mining such unprecedentedly large-scale and fine-granular spatial-temporal data. In this way, our research provides new approaches and insights into measuring economic activity.

Keywords: computational social science; economic activity; mobile data; complex systems

1 Introduction

The mobile internet, especially location-aware services, is ubiquitous in our everyday lives: each time a user opens an application (app), searches for a nearby restaurant, takes a car using a ride-hailing app, or uses mobile map navigation services, the user's location is detected via Global Positioning System (GPS) technology and logged on a server, generating massive mobility trace data. Such mobile big data, which directly reflect users' social and economic behaviours, provide new tools to measure economic dynamics in real time; thus, they overcome limitations on the timeliness and sample size of traditional survey methodologies and have deeply influenced empirical research [1–3].

China (the world's second-largest economy), with approximately 600 million smartphone users [4], has been profoundly affected by the mobile internet, even as it struggles to transform its economy from investment led to consumer driven. Thus, the indices used to gauge China's economic activities are valuable to researchers, investors, and most importantly, policy makers. However, there are three primary challenges to addressing this

data-driven issue. First, government statistics are generally released after a lag of weeks or months, and for some large-scale surveys (e.g., the economic census), coarse aggregate data are available only several years later. More timely indicators would undoubtedly help both governments and companies make more timely decisions. Second, many researchers question the reliability and quality of official data. For example, Gross Domestic Product (GDP) data are considered to be overstated [5], and the (registered) unemployment rate is suspected to be understated, as it remains steady at 4% and is hardly affected by economic slowdowns [6]. Moreover, many important economic indicators, such as surveyed unemployment, are not publicly available, casting a veil over the measurement of China's economic activity. Finally, but most importantly, the changing economic structure raises new challenges for measuring the emergence of service industries such as retail, restaurants, entertainment, and finance, which increasingly compose a considerable proportion of the economy but have been difficult to quantify.

Facing these challenges, researchers have previously turned to new data sources, such as search queries [7–16], social media [17–20], satellite images [21–25], online commodity pricing [26, 27], financial transactions [28–31], check-in data [32] and mobile phone data [33–40], to build socio-economic indicators or to study economic behaviours from different perspectives.

Varian *et al.* demonstrate the possibility of using Google search query indices for short-term economic prediction [9–11]. Additionally, web search data have proven to be helpful in forecasting consumer behaviour [12] and even financial market activity [13–15]. A similar methodology has been applied to social media; researchers have used Twitter data to create indices to predict economic activity such as unemployment [17, 18], stock market activity [19], and box-office earnings [20].

However, the analysis of search queries or social media texts is open to contestation, especially when terms with different meanings are employed [10]. Additionally, these works rely on 'ground truth' or official statistics as baseline models and then improve forecasting performance by utilizing search query or social media text data.

Another line of research seeks to identify proxy variables to measure economic activity directly, especially for countries for which official statistics are not available or reliable. For example, night-time light and remote sensing data are widely used to measure economic output [21–25] and poverty [41, 42]. Similarly, online product pricing data are scoured by researchers constructing online price indices, which not only are able to closely reflect government figures in countries such as the United States but also are useful for countries lacking trustworthy statistics [26, 27]. In industry, SpaceKnow and Baidu use satellite images and online web data to create the Satellite Manufacturing Index [25] and the Small Businesses Indices [43] in order to gauge China's economic activities.

Recently, academia and industry have begun using increasingly more individual-level data (especially financial transaction and mobile phone data) to study economic activity. For instance, Gelman *et al.* [28] measure people's spending responses to anticipated income using the transaction data of 75,000 users in the United States, as captured by Check (a financial service application). Agarwal and Qian [29] examine consumer responses to unanticipated income shocks using the bank transactions of 180,000 individuals in Singapore. China UnionPay has also developed economic indicators based on bankcard expenditures in various market segments [31]. Toole *et al.* [35] and Almaatouq *et al.* [36] track employment shocks using mobile phone Call Detail Records (CDRs) in Europe and Saudi

Arabia, respectively. Blumenstock *et al.* [37] infer poverty and wealth status at the individual level by combining Rwanda's mobile phone metadata and survey data with machine learning algorithms. Pappalardo *et al.* [38] propose a data-driven analytical framework to 'nowcast' socio-economical indicators using mobility features extracted from mobile phone data.

Despite being quite useful, mobile phone CDRs or metadata are still an indirect way to measure the economy; thus, they need to be combined with survey data or official statistics to inform models. Further, the spatial resolution of mobile phone data (at the cell tower level) ranges from hundreds of metres in the urban core to thousands of metres in rural areas, making it difficult to track firm-level economic activity.

The prevalence of smartphones (and the geo-located data and mobility trace data they generate) allows us to measure economic activity in China in a much more direct fashion and at a more granular level than has been possible using other previously explored data sources. In this paper, we build an *Employment Index* and a *Consumer Index* to measure employment trends in industrial parks and consumer activity in commercial areas by using billions of geo-positioning points (Figure 1). We evaluate the consumer index at the firm level by comparison with revenue data. Using location search data derived from Baidu Maps, we then propose models to estimate consumer foot traffic volumes for offline stores. We first apply these models to predict revenues of Apple retail stores and box-office earnings in China, and we achieve satisfactory results. Finally, we construct *Consumption Trends* to track consumer spending trends in various service sector industries (e.g., auto sales, restaurants, financial investments, and tourism) and verify them with several existing indicators.

To the best of our knowledge, this is the first study to measure the world's second-largest economy by mining such unprecedentedly large-scale and fine-granular spatial-temporal data. Our research, which provides new insights into China's economy, is designed not to

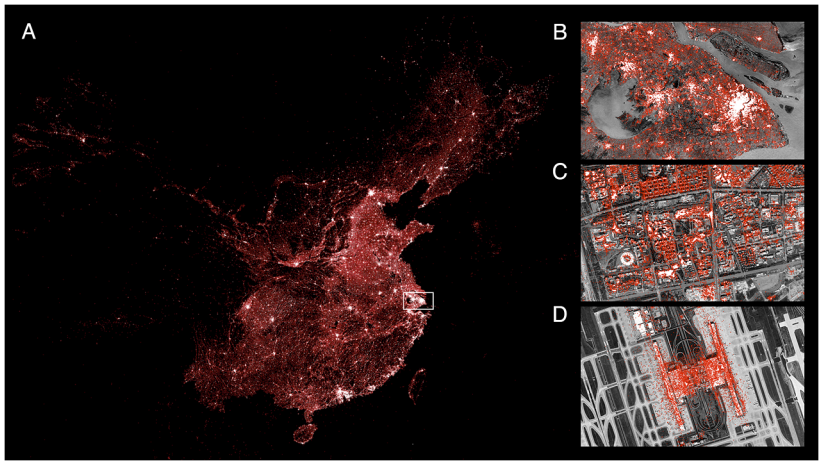

Figure 1 Spatial-temporal big data reflects human activity at different scales. (A) At the national level: Data points depict the fact that most of China's population is concentrated in large cities in the east. As the brightness of the spot increases, the aggregation of the data points (population) increases. **(B)** At the regional level: This figure shows urban clusters in the Yangtze River Delta. **(C)** At the zone level: Zhangjiang Hi-tech Park in Shanghai. **(D)** At the building level: Pudong Airport in Shanghai. Maps were created using C and Datamaps (https://github.com/ericfischer/datamaps), and the remote sensing images were derived from Baidu Maps.

supplant traditional surveys but to supplement such indicators in order to achieve more complete measurements.

2 Data

We use four datasets in this study:

(1) Geo-positioning data. Baidu Maps provides location search and positioning services for hundreds of millions of users, generating tens of billions of location requests each day. Each location point includes an anonymized ID, coordinates (longitude and latitude) and a timestamp. We gather geo-positioning data from January 2014 to June 2016. Figure 1 shows the spatial distribution of the positioning data.

(2) Location search data from Baidu Maps. Each location search datum (i.e., map query) includes an anonymized ID, query keywords, the returned Point of Interest (POI) ID, and a timestamp. As location search behaviour is strongly correlated with users' actual visits to the queried POI, we use online map query data to estimate offline foot traffic. We gather map query data for the period between January 2014 and June 2016.

(3) POI data, corresponding to a specific location such as a restaurant, hotel, or shopping mall. The whole dataset comprises approximately 50 million POI, all of which are classified into different categories by machine learning algorithms. For example, 'Walmart' belongs to the 'Supermarket' category, which is a subcategory of 'Shopping' (POI categories are detailed in Additional file 1). A map query action can be regarded as a consumer demand or as a future check-in for a certain type of place.

(4) Aera of Interest (AOI) data, corresponding to a specific type of region, such as an industrial park, commercial area, or scenic region. To measure employment and consumer economic activity, we manually label approximately 6,000 AOI, including 2,000 large industrial parks and 4,000 commercial areas. Supplementary Figure S1 (in Additional file 1) shows the spatial distribution of the AOI.

3 Methods

3.1 Sampling

In line with traditional survey practices that carefully draw samples based on various criteria, we sample mobile users who consistently registered location points during a 13-month rolling window (at least one point in each month). By using this method, we are able to draw a stable and sufficiently large sample of users and to reduce potential biases generated by changes in online services. The 13-month window also provides a convenient way to calculate year-over-year changes in figures.

3.2 Work place detection

We calculate the work location of each user via the following steps: (1) Filtering mobility traces: We extract positioning data during work time, which is defined to last from 9:00 a.m. to 6:00 p.m., excluding weekends and public holidays. (2) Clustering: We then adopt the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm [44] to calculate the work point of each user. DBSCAN is a density-based clustering algorithm, and it is commonly used to cluster trajectory data because of its accuracy and efficiency [45]. Two parameters are required for clustering: ε, the maximum distance between points to form a cluster; *MinPts*, the minimum number of points needed to form a

cluster. We set ε as 500 m and *MinPts* as 3 (see Supplementary Figure S2 for the robustness check of parameters). (3) Determining users' work place: We choose the centre of the cluster with the highest number of points as the user's work place.

3.3 Consumer detection

To track changes in the number of consumers, we sample users possessing continuous positioning points during a 13-month rolling window. We then count their monthly visits to the labelled commercial AOI. To reduce the possible influence of passers-by, only users who stay for a time period longer than a minimal threshold (10 minutes in our analysis, see Supplementary Figure S3 for the robustness check) in one day are identified as consumers.

3.4 Map query pre-processing

We apply map query data at the firm and sector levels to infer which consumers paid visits. We first choose the POI associated with particular economic activities according to their categories. Under the assumption that map query actions can be regarded as future check-ins at certain places [46, 47], we use the volumes of map queries to represent the number of potential consumers. For each map user, we drop duplicate queries within the same hour to reduce noise. Then, we count map queries for POI and aggregate them according to their categories. Finally, we scale map queries for each category by total query volumes to reduce the effect of user growth trends.

4 Results

4.1 Employment and consumer indices

Employment and consumption are two major indicators of economic activity. At the micro level, changes in the number of employees may reflect the performance of one specific company: If the business of a given company grows, more jobs will be created to expand production or services. By contrast, the pace of layoffs may accelerate when that company faces difficulties in business operations or a weakening market. These phenomena are closely bound to the macro-economy. For example, when companies suffered serious crises in the beginning of 2009, China's economy also experienced a rapid decline, with exports stumbling and demand contracting. This situation led to ever-increasing pressure from growing unemployment.

4.1.1 Tracking employment changes through geo-positioning data

We assume that the numerical changes of employees and consumers could be tracked from geo-positioning data, by which we construct indices to measure economic activity. In order to verify this idea, we study four cases with obvious employment volatility: two announced mass layoffs, and the remaining two experienced rapid developments with an employment surge. We first identify employees and count the monthly number in these areas by method mentioned in Section 3.2.

Figures 2A-C show the first two cases. Firm I is a large shoe factory located in a south-eastern coastal city in Guangdong Province, the heart of China's export manufacturing. Because of global industrial transfers, increasing labour costs and reduced export orders, the plant was closed in the first quarter of 2016, leaving thousands of workers jobless. As shown in Figure 2C, there was a sharp drop in the number of employees (the dark blue line) during this period (the grey box). A similar shock hit Firm II, a mobile phone factory

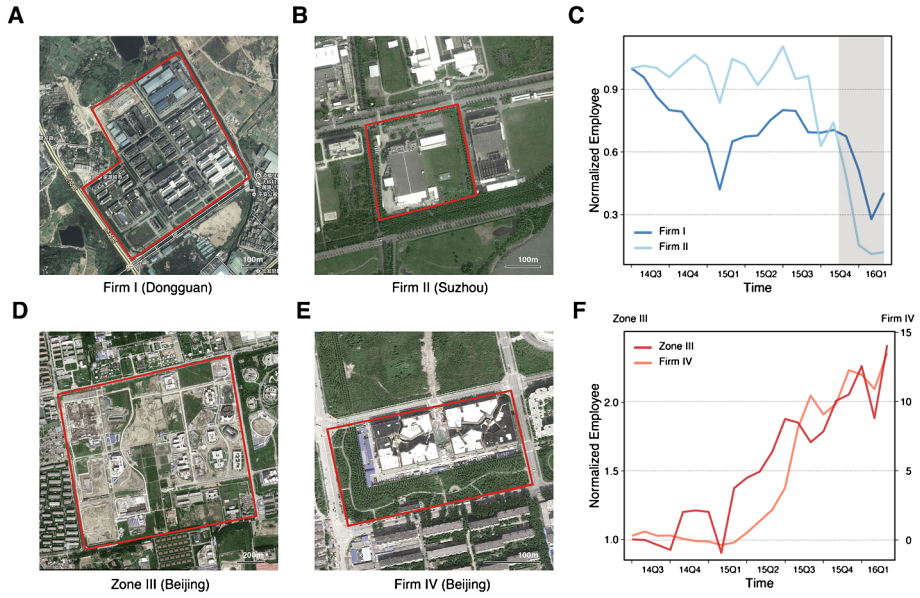

Figure 2 Measuring employment changes at the company/zone level. We study four typical cases, Firm I **(A)** and Firm II **(B)** announced mass layoffs, and Zone III **(D)** and Firm IV **(E)** experienced rapid growth. The normalized employee data are shown in **(C)** and **(F)**. The closing times for Firms I and II are marked by the grey bar in **(C)**. Remote sensing images were derived from Baidu Maps.

located in Suzhou, Jiangsu Province (Figure 2B), when it was shut down at the end of December 2015, with hundreds of workers losing their jobs. It is worth noting that Firm I gradually laid off its employees over three years, whereas Firm II shut down suddenly at the end of 2015. The discrepancy between these two cases is also mirrored in our results, as shown in Figure 2C: Employment for Firm I declined 20% further in 2015 from 2014, with a consequent year-over-year decline of 30% in the first quarter of 2016. By contrast, employment for Firm II remained stable and then dipped dramatically in the last quarter of 2015, after the shutdown.

Figures 2D-F show the second two cases of increasing employment. Zone III is a software park in Beijing (Figure 2D), home to many high-tech companies. Most of these companies began to relocate there in 2014, and since then, as shown in Figure 2F (red line), the number of employees has almost doubled. Firm IV is a fast-growing start-up based in Beijing (Figure 2E) that has experienced rapid expansion since 2015. As the figure shows, its employment jumped in the second quarter of 2015 after it raised billions of dollars in investment funding.

4.1.2 Employment index

We build an employment index to track macro trends in employment by aggregating employment changes in 2,000 labelled AOI, including 1,000 manufacturing industrial parks and 1,000 high-tech industrial parks. Since we sample users with consistent location points during a 13-month rolling window, the first 13-months' index is simply the normalized number of workers in AOIs. Starting from the 14th month, the employment index is calculated by the following equations:

$$rate_t = \frac{\sum_i^n AOI_{i,t} - \sum_i^n AOI_{i,(t-12)}}{\sum_i^n AOI_{i,(t-12)}}, \tag{1}$$

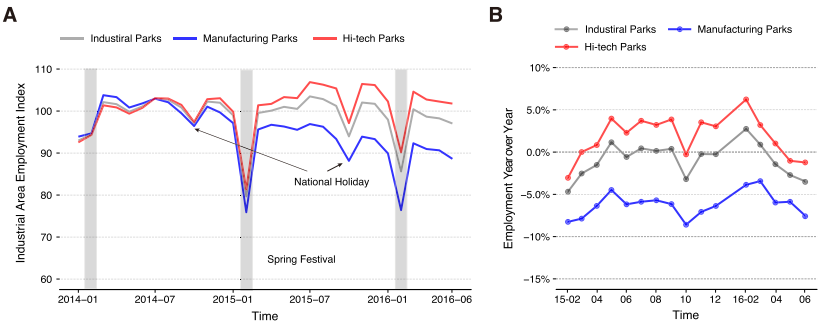

Figure 3 Employment index for industrial parks. (A) The grey line shows the overall trend among all 2,000 industrial parks, which is relatively steady with a slow decline. The blue line, which represents the result for approximately 1,000 manufacturing industrial parks, shows a sustained and more obvious downward trend; the red line, which represents the result for approximately 1,000 high-tech parks, shows healthy growth. **(B)** Monthly year-over-year growth rate. January and February are aggregated together to remove the influence of the Spring Festival holiday, which falls in a different time period in January or February every year.

$$employment_t = employment_{t-12} \times (1 + rate_t), \tag{2}$$

where $rate_t$ is the year-over-year percentage change of number of workers within all AOIs, and $employment_t$ is the employment index of month t.

Figure 3A illustrates the monthly employment index from January 2014 to June 2016 without seasonal adjustment, in which the grey, blue and red lines indicate the employment index for all industrial parks, manufacturing parks, and high-tech parks, respectively. All indices are normalized by the mean value for the year 2014 (set to be 100). Figure 3B shows the corresponding year-over-year growth rate. The following findings can be drawn from the figures: (1) The overall employment index of industrial parks was relatively steady, with a slow decline over the last two years; employment began falling in the beginning of 2016, indicating a general slowdown in the economy. (2) The monthly employment index of manufacturing parks showed a sustained and more obvious decline in overall performance; this downward trend is similar to what was found in the official manufacturing purchasing managers index (see Supplementary Figure S4). The employment index of manufacturing parks dipped more than 5% year-over-year during most of the study period, indicating that this sector seriously suffered from the economic slowdown. (3) The employment index of high-tech parks, by contrast, saw a healthy and continuous increase, with more than a 3.5% year-over-year increase over the last two years, which suggests that high-tech companies in industries such as software and bio-tech (referred to as belonging to the 'new economy') continued to create more jobs than other sectors.

4.1.3 Consumer index

In a similar fashion, we track changes in consumer activity for approximately 4,000 major commercial areas, covering most of the large shopping malls in China. We then construct a monthly consumer index to measure the changes and track the growth in consumer activity in these areas using Eq. 1 and Eq. 2. The consumer index generally climbs during the summer holidays (July and August) and before the Spring Festival in December and January and then declines during the Spring Festival in February (Figure 4A). One can see that the consumer index gradually inches down and that its year-over-year growth rate fluctuates, with more dips below zero than spikes over time, indicating weakening

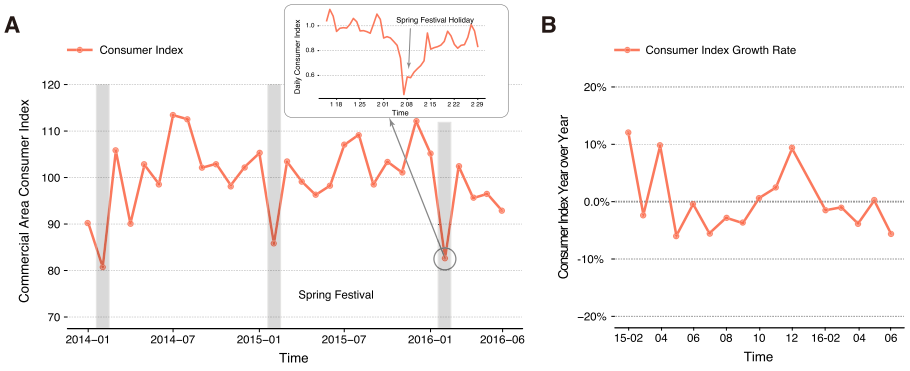

Figure 4 Consumer index for commercial areas. (A) The orange line is the monthly consumer index, which is normalized by the mean value of the year 2014, and the Spring Festival is marked by the grey bar. **(B)** Monthly year-over-year growth rate. The consumer index generally climbed during the summer holiday (July and August) and before the Spring Festival (December and January) and decreased during the Spring Festival (February). The year-over-year growth rate fluctuated, with more dips below zero than spikes over time, indicating that the consumer index trend decreased.

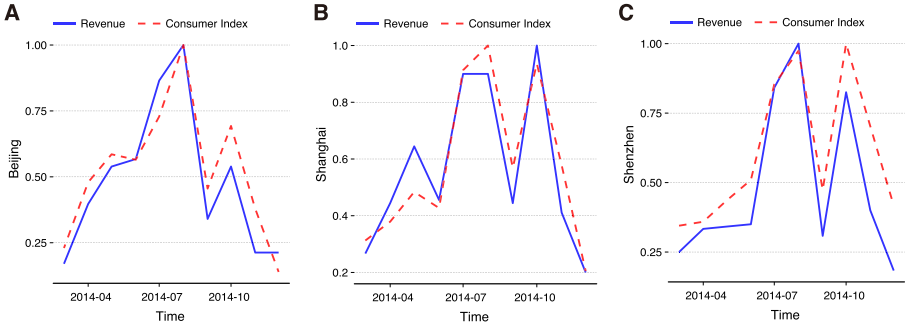

Figure 5 Consumer index and revenue. (A) Commercial area I at Beijing. **(B)** Commercial area II at Shanghai. **(C)** Commercial area III at Shenzhen.

consumer demand (possibly dragged down by the overall slowdown and by e-commerce). The year-over-year growth rate shown by the consumer index (Figure 4B) began rising in September 2015 and peaked at approximately 9% in December; however, at the beginning of 2016, it fell back below zero and sank to nearly –7% in June of this year, indicating relatively low consumption demand in the first half of 2016.

To investigate whether consumer index matters for economic performance, we evaluate our index by comparison with firm level revenue. We collect monthly revenue data of three large commercial areas, located in Beijing, Shanghai, and Shenzhen respectively. For comparison, we rescale the consumer index of each commercial area, and plot them together with corresponding revenue data. Figure 5 shows that our consumer index is highly correlated with the revenue data, and the Pearson correlation coefficient is 0.94 for each city, suggesting that our consumer index could be a reasonable proxy for measures of commercial performance.

In summary, the proposed employment index and consumer index measure the economic performance of selected industrial parks and commercial areas in China, depict the complicated reality of China's economy, and provide a new perspective for charting macro-economic performance from a bottom-up view.

4.2 Foot traffic estimation, revenue forecast, and consumption trends

In the previous section, we construct the consumer index by counting the number of consumers who visited large commercial areas according to the geo-positioning dataset. In this section, we seek to track consumer spending trends in various service sector industries. This task, however, is challenging because using only geo-location data to identify users' offline visits to specific POI such as a restaurant or retail shop rather than to an AOI such as a shopping centre cannot achieve ideal accuracy. Although the computer science community has proposed several machine learning models designed to relate specific POI to mobility traces (this challenge is referred to as the 'trajectory semantic problem') and has applied these models to several business scenarios, even the state-of-the art models [48] are not sufficiently accurate for rigorous social and economic science research. We turn to a new data source, location search data from Baidu Maps (i.e., map query data), to address the following questions: (1) Is it possible to estimate offline foot traffic for one specific location? (2) Can we 'nowcast' or even forecast the consumer spending or revenue for one location or for a firm with chain stores? (3) How can we build consumer spending indices to evaluate the performance of different industries?

4.2.1 Foot traffic estimation from map query data

Foot traffic is one of the most critical metrics for service sector businesses such as retail stores, restaurants, movie theatres, and hotels. We observe that, unsurprisingly, the volume of map queries regarding one specific location is highly correlated with that of offline foot traffic. This correlation is illustrated in Figure 6, which shows the number of people searching for Shopping Malls I and II (orange dashed line) on Baidu Maps and the

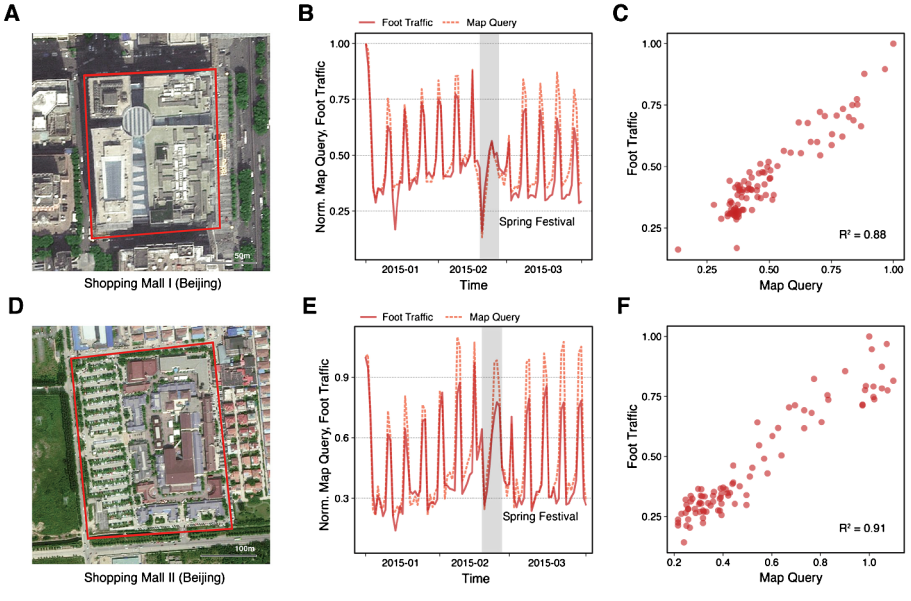

Figure 6 Map queries and foot traffic. Spatial layout of Shopping Malls I **(A)** and II **(D)**. The number of people searching for Shopping Malls I **(B)** and II **(E)** on Baidu Maps and foot traffic for each place. Scatter plots of map queries and foot traffic for Shopping Malls I **(C)** and II **(F)**. In both cases, the map query data were highly correlated with the offline foot traffic data, with both showing cyclical patterns in which they peaked on weekends. Maps were created using Adobe Illustrator, and the remote sensing images were derived from Baidu Maps.

Table 1 Regressions of foot traffic and map queries

Variables	Shopping Mall I (1)	Shopping Mall I (2)	Shopping Mall II (1)	Shopping Mall II (2)
y_{t-1}	0.125	0.117**	0.072	0.080*
	(0.068)	(0.038)	(0.062)	(0.034)
y_{t-7}	0.811***	0.078	0.866***	0.030
	(0.068)	(0.072)	(0.062)	(0.076)
Map Query		2.60***		5.10***
		(0.220)		(0.413)
Observations	74	74	74	74
R-squared	0.727	0.914	0.767	0.931

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

foot traffic (red solid line) of each place (estimated from the geo-positioning dataset). In both cases, map queries and foot traffic present obvious periodical patterns, peaking on weekends and dipping during weekdays. To evaluate the efficiency and accuracy of the foot traffic estimation via map query data, we compare two seasonal autoregression (AR) models to predict foot traffic:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-7} + e_t, \tag{3}$$

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-7} + \beta_3 q_t + e_t, \tag{4}$$

where y_t, y_{t-1}, and y_{t-7} are the numbers of consumers at time t, $t-1$, and $t-7$, respectively. Here we add y_{t-7} to the regression because human activities have weekly periodicity. The volume of map queries at time t is q_t, and e_t is the error term. The regression results, shown in Table 1, indicate that adding map queries to the regression significantly improves the in-sample fit. The R^2s of Shopping Malls I and II increase from 0.727 and 0.767 to 0.914 and 0.931, respectively. In fact, if we use only map queries to fit the foot traffic estimation, the results are surprisingly accurate: the R^2s are 0.880 for Shopping Mall I and 0.910 for Shopping Mall II (Figure 6CF). Therefore, we can use map queries to estimate the volume of foot traffic and evaluate the performance of foot-traffic-dominated stores, firms, and industries in the service sector.

4.2.2 Foot traffic and revenue forecast

Here, we demonstrate two cases to validate the potential power of map queries for estimating and forecasting revenue. The first one is a revenue prediction for Apple Inc. (Apple) in China, and the second is box office 'nowcasting' and fraud detection.

4.2.3 Apple sales in China

As a listed company, Apple reports its earnings results to the public quarterly with approximately one month's lag time. According to official numbers, Apple reported $75.9b in revenue during the last quarter of 2015 (note that the fiscal quarter is different from the calendar quarter), and China, the second-largest market globally, contributed $18.4b (in the same quarter a year prior, the total revenue was $74.6b, and the Chinese market contributed $16.1b). In the first quarter of 2016, Apple's revenue ($50.6b) was down year over year for the first time since 2003, and its revenue in China declined by approximately 26%.

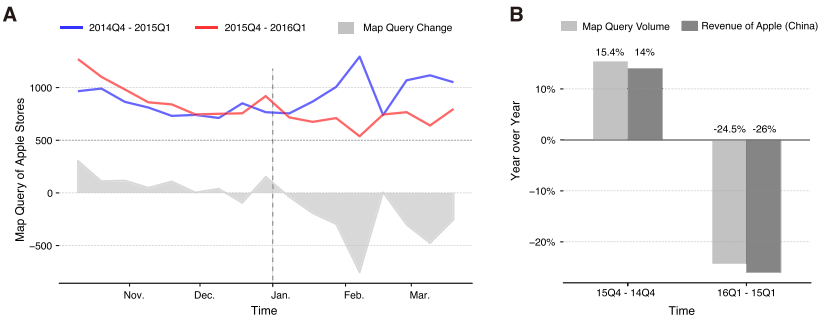

Figure 7 Map queries for Apple Stores and revenues. (A) The volume of map queries for flagship Apple Stores in Mainland China. Year-over-year changes in queries are plotted as grey areas. **(B)** The volume of map queries showed a 15.4% year-over-year increase in the fourth quarter of 2015 and a 24.5% decline in the first quarter of 2016, and revenues showed a 14% increase and 26% decrease during the corresponding period.

Given that we are able to estimate foot traffic volumes for offline stores by using map queries, should we conclude that these volumes are related to reported revenue? Is it possible for us to predict quarterly revenue prior to earnings announcements?

To answer these questions, we first select a list of flagship Apple Stores in Mainland China (see Additional file 1 for the list) and then count the volume of map queries for all the stores. As shown in Figure 7A, the blue line spans from the last quarter of 2014 to the first quarter of 2015; the red line spans the corresponding time period for 2016. We also calculate the year-over-year changes in map queries (the grey area in Figure 7A) and compare it with revenues (Figure 7B). As Figure 7 shows, the volume of map queries showed a 15.4% year-over-year increase in the last quarter of 2015 and a 24.5% decline in the first quarter of 2016, and revenues similarly showed a 14% increase and a 26% decrease during the corresponding period. The impressively strong correlation indicates that map query data allow us to 'nowcast' the company's revenues and reveal future trends. We then applied this method to map query data of the second quarter of 2016 and projected that Apple's revenue in Greater China in this quarter may decline 34% on a year-over-year basis (one month prior to earnings announcements). The final number in the earnings reports was 33%, which is close to our prediction.

4.2.4 Movie box office 'nowcasting' and fraud detection

Going beyond estimation at the firm level, we find that map query data are also predictive at the level of the specific consumer sector. To better illustrate, we collect daily box office data during 2015 from the China Box Office Database [58921.com] and then build models to 'nowcast' daily box office revenues by using map queries related to cinemas. Here, we use the word 'nowcasting' instead of 'forecasting' because we use map query data to predict box-office revenues for the same day, which is approximately one day earlier than the official statistics. To reduce the effect of user search trends, we normalize the volume of daily map query data and use it as the independent variable in our regression models. Our baseline model is a simple AR model with 1-day and 7-day time lags, as we have no further information about movies (e.g., production budgets and number of opened screens, which are used as input features in ref. [12]). The regression equations are similar to Eq. 3 and Eq. 4.

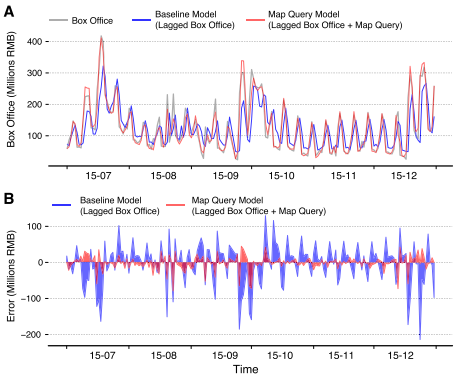

Figure 8 Box office prediction. (A) Box office revenue (grey line), baseline model (blue line) and map query model (red line). **(B)** Prediction errors of the baseline model (blue line) and the map query model (red line). The absolute error of the baseline model was almost always larger than that of the map query model, and the MAE of the map query model declined 67% relative to the baseline model.

As shown in Supplementary Table 1, map queries for cinemas significantly improve the in-sample fit: the R^2 of the baseline model is 0.489, while that of the map query model is 0.934. To further investigate the models' out-of-sample performance, we apply a rolling window forecast similar to the method used in [10, 49]. Given the selected date (July 1st, 2015, in this case), prediction in the $(t-1)$ time step is based on estimates of all previous time periods ($\leq t$).

The results are plotted in Figure 8, where the grey line represents the actual box office statistics, the blue line represents the forecasting results of the baseline model, and the red line denotes the forecasting results of our map query model. The absolute error of the baseline model is almost always larger than that of the map query model. The mean absolute error (MAE) of the map query model declines 67% below that of the baseline model, with a value of 13.2 million RMB (the mean of daily box office is 118 million, with a standard deviation of 75.8 million), as shown in Figure 8B.

Further, we reveal some additional insights when analysing data for specific movies and find that map queries are helpful for detecting box office fraud. For example, *Monster Hunter*, a Chinese box office champ in 2015 earning approximately 2.4 billion RMB in revenue, was reported to have artificially inflated its box office by scheduling midnight screenings with almost no audience present in cinemas owned by the movie's producer [50]. To check for such inflated numbers, we select a list of suspected cinemas owned by the producer of *Monster Hunter* (approximately 30 cinemas) and a list of 30 other cinemas as a control group. We then calculate the volume of map queries for these locations from July 2015 to September 2015, when the movie was shown.

Figure 9A shows the results of suspected cinemas, and Figure 9B shows those of the control group. The grey lines and red lines indicate the total box office revenues and *Monster Hunter* box office revenues, respectively. The blue lines show the fitting results by map queries, and the black dashed lines denote forecasting errors. Comparing the forecasting errors for two sets of selected cinemas, one can see that the forecasting error for the control group is small and stable, whereas for the suspected cinema set, there is a sharp increase in the forecasting error from August 25 to September 16 (marked by the grey box). Such an anomaly indicates a high probability of box office fraud.

4.2.5 Consumption trends

As demonstrated above, the volume of map queries is strongly correlated with foot traffic and is able to estimate offline consumer spending. We therefore construct the consump-

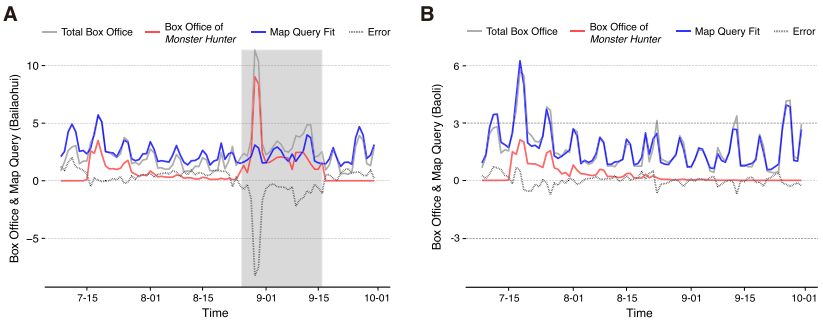

Figure 9 Box office fraud detection. We selected 30 POI of suspected cinemas **(A)** and another 30 cinema POI as a control group **(B)**. The forecasting error for the control group was small and stable, whereas for the suspected cinema set, there was a sharp increase in the forecasting error from August 25 to September 16 (marked by the grey box). Such an anomaly indicates a high probability of box office fraud.

Figure 10 Consumption trends for various industries in the service sector. All indices are normalized by the mean value for the year 2014. **(A)** Auto sales. **(B)** Restaurants. **(C)** Finance & Investments. We mark the 'turning point' in stock market and map query data with circles, which indicate that map queries may be a leading indicator for stock market movements. **(D)** Tourism. Important holidays are labelled with arrows.

tion trends for various industries by aggregating the map queries belonging to the same POI category. For example, one who searches for a car dealership on a map will be regarded as a potential automobile industry consumer. In the main text, we present consumption trends for four sectors, *Automobile Sales*, *Restaurants*, *Finance & Investments*, and *Tourism*, and compare our results with several other indicators (see Supplementary Figures S5-S7 for more sectors' trends). All trends are normalized by the mean value for the year 2014.

Figure 10A shows the overall rising trend in automobile sales based on an analysis of approximately 190,000 related locations, which is highly correlated with the sales data

published by the Automobile Association of China [http://www.caam.org.cn/]. The Pearson correlation is 0.913, with $p < 0.001$, indicating that the proposed trend can effectively track market changes in real time.

Figure 10B shows the relative declining trend in the restaurant sector, which includes almost 3 million restaurants. We compare our result with the Tsinghua UnionPay Advisors' Restaurant & Catering (TUPARC) Index [31], a consumer spending indicator based on bank card transaction data from the China Cuisine Association's 100 top-ranked restaurants. While our proposed trend is consistent with the TUPARC over most of the time period (the Pearson correlation coefficient is 0.764, with $p < 0.001$), our restaurant consumption trend began dipping in 2016 and fell to a new low of 85 in June.

Figure 10C shows results for the finance and investment sector, which includes 230,000 related locations. Interestingly, we find that our proposed trend resembles the trends of the Shanghai Composite Index, one of the stock market indices in China. Our on-going investigation indicates that the proposed trend may be a leading indicator for stock market movements, especially for 'turning points' (marked with circles). Specifically, we follow the trading strategy proposed in [13] to analysis weekly changes in map query volumes for investment sector (e.g. banks, security companies). If the query volume of week t (note as q_t) overs the average volume of three weeks before, we would take the 'long position': buy the stock index at the closing price on the first trading day of week $t+1$ and sell at the closing price on the last trading day of week $t+1$. Otherwise, we would take the 'short position': sell at the closing price of the first trading day and buy at the last trading day of week $t+1$:

$$\text{if} \quad q_t > \frac{q_{t-1} + q_{t-2} + q_{t-3}}{3} : \quad \text{Long position,}$$

$$\text{elif} \quad q_t \leq \frac{q_{t-1} + q_{t-2} + q_{t-3}}{3} : \quad \text{Short position.}$$

We apply this weekly strategy to the stock market dataset in the period between Jan. 2014-Apr. 2016 (120 weeks). If we buy at the beginning week (the Shanghai Composite Index is 2038) and hold until the end (the Index is 2956), the yield is approximate 17.5% annually. In contrast, if we take the 'long/short strategy' as mentioned above, the yield is approximate 49.3% annually (transaction fees are neglected in this analysis).

Figure 10D shows the trend for the travel sector, which comprises 300,000 tourist attractions. The overall trend grows moderately, exhibiting obvious spikes during major Chinese holidays, such as the Spring Festival, the National Holiday and the summer holiday, which generally lasts from July to August.

Our consumption trends for these sectors demonstrate that, on a year-over-year basis, automobile sales rose, restaurant spending declined, finance investment was related to stock market volatility, and travel spending grew moderately. All these findings reflect the complicated reality of China's service economy.

5 Discussions

In summary, we use mobile big data to construct real-time indices to track trends in Chinese employment and consumption. We employ several strategies to verify our assumptions and results. At the national level, because there are no parallel statistics in government reports, we study several cases covered by the media and find that our data and

method can successfully track changes in the number of employees in these cases. At the firm and sector levels, we estimate foot traffic by using map queries, and we obtain impressive results in predicting Apple's revenues and box-office earnings in China. We also construct indices to track consumer spending trends for various service-sector industries and compare our results with other existing indicators.

Using mobile data to build economic indicators has great practical value. Compared with data obtained from traditional survey-based methods, mobile data are available in almost real time, offering larger sample sizes, finer resolution, and lower costs. Further, compared with social media data, mobile data boast a higher coverage rate, and they are more structured and robust [35]. These features are quite valuable to market participants and policy makers, who need timely and reliable data to make decisions. More importantly, the nature of mobility traces and location search data allows us to measure individual-level behaviours in a much more direct fashion. Our paper belongs to the research concept of *Mobimetrics*, which dedicated to quantifying social system dynamics by analysing massive individual mobility data generated by smartphones, wearable devices, driverless cars, and even the Internet of Things in the near future with machine learning approaches [32–40]. These mobile data, combined with machine learning methods, have the potential to change the landscape of economics and social science empirical research.

Although we highlight the value of mobile data in measuring and predicting economic activity, we do not claim that it can supplant official statistics. In particular, mobile data cover only people who use specific services, introducing potential bias to topic-specific research [51]. For example, the percentage of elderly people and children using the devices and apps that generate mobile data is relatively low, so the data do not reflect China's real age structure (a comparison of the users' ages and location distributions versus the survey results is shown in Supplementary Figure S8). Thus, if a study is designed to evaluate a policy affecting the elderly population, mobile data may not be a suitable choice. In fact, the combination of survey and new data sources may offer greater benefits as a result of their respective advantages, allowing researchers to achieve more accurate and complete measurements.

This study marks only the beginning of the application of *Mobimetrics*, and several improvements could be made in future research. First, to cover more sectors and to build a more comprehensive economic measure, more AOI could be automatically labelled by using machine learning and incorporating more data sources, such as road network data and satellite images. Second, the proposed economic indices could be further investigated through cross-validation with additional data sources. For example, map queries for hospitals, which can be viewed as an important indicator for disease monitoring, could be verified by using data from the Centres for Disease Control and Prevention. Third, the on-demand platform economy, or the sharing economy, is now playing an increasingly important role in our daily lives, and identifying methods for measuring such economic activity via mobile data is worth exploring.

Additional material

Additional file 1: Supplementary Material for Measuring Economic Activity in China with Mobile Big Data.
(pdf)

Acknowledgements
We thank all the members of the Spatial Temporal Big Data Group of Big Data Lab (BDL) for a helpful discussion.

Funding
This research was supported by the National Natural Science Foundation of China (No. 41625003), and National Science and Technology Major Project (No. 2017YFB0503602).

Abbreviations
GPS, Global Positioning System; GDP, Gross Domestic Product; CDRs, Call Detail Records; POI, Point of Interest; AOI, Aera of Interest; DBSCAN, Density-Based Spatial Clustering of Applications with Noise; AR, Autoregression; MAE, Mean Absolute Error; TUPARC, Tsinghua UnionPay Advisors' Restaurant & Catering Index.

Ethics approval and consent to participate
We adopted a very rigorous protocol in this research to protect users' privacy: (1) All user IDs in our data were hashed and anonymized to ensure that one could not associate data with individual users. (2) All the data were saved in secure servers, and they could be accessed only by following strict procedures with high standards. (3) All the researchers were required to adhere to a confidentiality agreement, which required them to use data only for approved research. In this research, we focus solely on aggregated data instead of individual-level data to measure economic activity.

Competing interests
HW, YC and CL are full-time employees of Baidu. LD, SC, and ZW were interns at Baidu during the research period. The authors declare that they have no competing interests.

Authors' contributions
LD and HW designed the research and wrote the paper. LD, SC, YC and ZW performed the data analysis. CL helped with data pre-processing. All authors reviewed the manuscript, read and approved the final manuscript.

Author details
[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.

Publisher's Note

Received: 2 January 2017 Accepted: 17 October 2017 Published online: 06 November 2017

References

1. Einav L, Levin J (2014) Economics in the age of big data. Science 346(6210):1243089
2. Varian H (2014) Big data: new tricks for econometrics. J Econ Perspect 28(2):3-27
3. Lazer D, Pentland A, Adamic L, Aral S, Barabasi AL, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Alstyne M (2009) Computational social science. Science 323(5915):721-723
4. CNNIC (2016) China internet development report. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/201601/P020160122469130059846.pdf. Accessed 13 Oct 2016
5. Rawski TG (2001) What is happening to China's GDP statistics? China Econ Rev 12(4):347-354
6. Feng S, Hu Y, Moffitt R (2015) Long run trends in unemployment and labor force participation in China. NBER Working Paper No. 21460
7. Ettredge M, Gerdes J, Karuga G (2005) Using web-based search data to predict macroeconomic statistics. Commun ACM 48(11):87-92
8. Askitas N, Zimmermann KF (2009) Google econometrics and unemployment forecasting. Appl Econ Q 55(2):107-120
9. Choi H, Varian H (2009) Predicting initial claims for unemployment benefits. Google Technical Report 1-5
10. Choi H, Varian H (2012) Predicting the present with Google trends. Econ Rec 88(s1):2-9
11. Scott SL, Varian H (2013) Bayesian variable selection for nowcasting economic time series. NBER Working Paper No. 19567
12. Goel S, Hofman JM, Lahaie S, Pennock DM, Watts DJ (2010) Predicting consumer behavior with web search. Proc Natl Acad Sci USA 107(41):17486-17490
13. Preis T, Moat HS, Stanley HE (2013) Quantifying trading behavior in financial markets using Google trends. Sci Rep 3:1684
14. Preis T, Moat HS, Stanley HE, Bishop SR (2012) Quantifying the advantage of looking forward. Sci Rep 2:350
15. Curme C, Preis T, Stanley HE, Moat HS (2014) Quantifying the semantics of search behaviour before stock market moves. Proc Natl Acad Sci USA 111(32):1160-11605
16. Yang X, Pan B, Evans JA, Lv B (2015) Forecasting Chinese tourist volume with search engine data. Tour Manag 46:386-397
17. Antenucci D, Cafarella M, Levenstein M, Ré C, Shapiro MD (2014) Using social media to measure labor market flows. NBER Working Paper No. 20010
18. Llorente A, Garcia-Herranz M, Cebrian M, Moro E (2015) Social media fingerprints of unemployment. PLoS ONE 10:e0128692
19. Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. J Comput Sci 2(1):1-8
20. Asur S, Huberman BA (2010) Predicting the future with social media. In: International conference on web intelligence and intelligent agent technology (WI-IAT), vol 1. IEEE/WIC/ACM, pp 492-499
21. Chen X, Nordhaus WD (2011) Using luminosity data as a proxy for economic statistics. Proc Natl Acad Sci USA 108(21):8589-8594

22. Henderson JV, Storeygard A, Weil DN (2012) Measuring economic growth from outer space. Am Econ Rev 102(2):994-1028
23. Michalopoulos S, Papaioannou E (2013) Pre-colonial ethnic institutions and contemporary African development. Econometrica 81(1):113-152
24. Mellander C, Lobo J, Stolarick K, Matheson Z (2015) Night-time light data: a good proxy measure for economic activity? PLoS ONE 10:e0139779
25. SpaceKnow (2016) China satellite manufacturing index. http://spaceknow.com. Accessed 13 Oct 2016
26. Cavallo A (2015) Scraped data and sticky prices. NBER Working Paper No. 21490
27. Cavallo A (2013) Online and official price indexes: measuring Argentina's inflation. J Monet Econ 60(2):152-165
28. Gelman M, Kariv S, Shapiro MD, Silverman D, Tadelis S (2014) Harnessing naturally occurring data to measure the response of spending to income. Science 345(6193):212-215
29. Agarwal S, Qian W (2014) Consumption and debt response to unanticipated income shocks: evidence from a natural experiment in Singapore. Am Econ Rev 104(12):4205-4230
30. JPMorgan Chase Institute (2016) Paychecks, paydays, and the online platform economy. https://www.jpmorganchase.com/corporate/institute/document/jpmc-institute-volatility-2-report.pdf. Accessed 13 Oct 2016
31. UnionPay (2016) UPA indices. http://www.unionpayadvisors.com/big_data_index.php. Accessed 13 Oct 2016
32. Glueck J (2016) Foursquare predicts Chipotle's Q1 sales down nearly 30%. https://medium.com/foursquare-direct/foursquare-predicts-chipotle-s-q1-sales-down-nearly-30-foot-traffic-reveals-the-start-of-a-mixed-78515b2389af. Accessed 13 Oct 2016
33. Blondel V, Decuyper A, Krings G (2015) A survey of results on mobile phone datasets analysis. EPJ Data Sci 4(1):10
34. Deville P, Catherine L, Samuel M, Marius G, Forrest RS, Andrea EG, Vincent DB, Andrew JT (2014) Dynamic population mapping using mobile phone data. Proc Natl Acad Sci USA 111(45):15888-15893
35. Toole JL, Lin YR, Muehlegger E, Shoag D, Gonzalez MC, Lazer D (2015) Tracking employment shocks using mobile phone data. J R Soc Interface 12(107):20150185
36. Almaatouq A, Prieto-Castrillo F, Pentland A (2016) Mobile communication signatures of unemployment. In: International conference on social informatics. Springer, Berlin, pp 407-418
37. Pappalardo L, Vanhoof M, Gabrielli L, Smoreda Z, Pedreschi D, Giannotti F (2016) An analytical framework to nowcast well-being using mobile phone data. Int J Data Sci Anal 2(1-2):75-92
38. Blumenstock J, Cadamuro G, On R (2015) Predicting poverty and wealth from mobile phone metadata. Science 350(6264):1073-1076
39. Marchetti S, Giusti C, Pratesi M, Salvati N, Giannotti F, Pedreschi D, Rinzivillo S, Pappalardo L, Gabrielli L (2015) Small area model-based estimators using big data sources. J Off Stat 31(2):263-281
40. Smith-Clarke C, Capra L (2016) Beyond the baseline: establishing the value in mobile phone based poverty estimates. In: Proceedings of the 25th international conference on world wide web. International World Wide Web Conferences Steering Committee, pp 425-434
41. Xie M, Jean N, Burke M, Lobell D, Ermon S (2015) Transfer learning from deep features for remote sensing and poverty mapping. arXiv:1510.00098
42. Jean N, Burke M, Xie M, Davis WM, Lobell DB, Ermon S (2016) Combining satellite imagery and machine learning to predict poverty. Science 353(6301):790-794
43. Baidu (2016) Small businesses indices. http://trends.baidu.com/economy/. Accessed Oct 13 2016
44. Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd international conference on knowledge discovery and data mining, pp 226-231
45. Borah B, Bhattacharyya DK (2004) An improved sampling-based DBSCAN for large spatial databases. In: Proceedings of the international conference on intelligent sensing and information processing, pp 92-96
46. Zhou J, Pei H, Wu H (2016) Early warning of human crowds based on query data from Baidu map: analysis based on shanghai stampede. arXiv:1603.06780
47. Xu M, Wang T, Wu Z, Zhou J, Li J, Wu H (2016) Store location selection via mining search query logs of Baidu maps. arXiv:1606.03662
48. Furletti B, Cintia P, Renso C, Spinsanti L (2013) Inferring human activities from GPS tracks. In: Proceedings of the 2nd ACM SIGKDD international workshop on urban computing. ACM, New York, 5
49. Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of Google flu: traps in big data analysis. Science 343(6176):1203-1205
50. Hoad P (2015) China box-office record breaker monster hunt 'manipulated figures'. https://www.theguardian.com/film/2015/dec/07/china-box-office-record-breaker-monster-hunt-manipulated-figures. Accessed Oct 13 2016
51. Ruths D, Pfeffer J (2014) Social media for large studies of behavior. Science 346(6213):1063-1064