Predicting socio-economic levels of urban regions via offline and online indicators

Yi Ren¹,², Tong Xia³, Yong Li³, Xiang Chen¹,²*

¹ School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China, ² Key Lab of EDA, Research Institute of Tsinghua University in Shenzhen (RITS), Shenzhen, China, ³ Department of Electronic Engineering, Tsinghua University, Beijing, China

* chenxiang@mail.sysu.edu.cn

Abstract

Predicting the socio-economic level of an urban region is of great significance for governments and city managers when allocating resources and making decisions. However, the current approaches for estimating regional socio-economic levels heavily rely on census data, which demands significant effort in terms of time and money. With the ubiquitous usage of smart phones and the prevalence of mobile applications, massive amounts of data are generated by mobile networks that record people’s behaviors. In this paper, we propose a low-cost approach of using humans’ online and offline indicators to predict the socio-economic levels of urban regions. The results show that the socio-economic prediction model that is trained using online and offline features extracted from these data achieves a high accuracy over 85%. Notably, online features are showed to be tightly linked with socio-economic development. In environments where censuses are rarely held, our method provides an option for timely and accurate prediction of the economic status of urban regions.

Introduction

The accurate and real-time prediction of the socio-economic level (SEL) of a region plays an important role in grasping the development level of the city’s region. The government makes its decisions regarding coordinating the overall development according to the distribution of SELs. The current methods to investigate SELs mainly rely on the economic census that is organized by the National Statistical Institute (NSI). The NSI organizes the economic census every five years in China, which makes it difficult to provide timely referential data. In addition, the NSI counts the Gross Domestic Product (GDP) every quarter, which incurs great manpower costs. Therefore, a relatively low-cost approach is needed to estimate the SELs of urban regions.

Studies about economics predictions [1–5] using novel datasets have been conducted for a few years. Previous works have used mobile phone data to evaluate socio-economic status [3, 6, 7]. Some studies [2, 8, 9] extracted social, behavioral and mobility features from calling details records and trained machine learning models based on these features. Luca Pappalardo et al. found that there is a tight correlation between the aggregated human mobility patterns...
that are discovered from mobile phone data and socio-economic indicators [10]. They computed the mobility volume and mobility diversity at the individual level and aggregated them at the municipality level. Ultimately, they indicated that the aggregated mobility measures are correlated with socio-economic indicators. However, due to the expansion of the Internet age, people use chat applications increasingly more frequently instead of making calls. Therefore, methods that use the features that are extracted from calling data to predict socio-economic levels may become less useful. Other researchers attempted to identify economic indicators using satellite imagery [11, 12] or night time light [4]. However, this prediction would be outdated due to the popularity of the infrastructure. In wealthy countries, the data that are collected from the Internet and social media have been used to evaluate SELs [13, 14]. For instance, studies have proposed approaches that utilize Google Trends to predict short-term values of economic indicators [15] and unemployment benefits [16]. Some other previous works showed that social networks [17–19] have greatly impacted economics.

Existing research concentrates on using the mobility features that are extracted from calling detail records to predict economic status. However, with the popularity of the mobile Internet and widespread usage of smart phones, applications are bleeding into daily life and calls are gradually being replaced. Thus, the online features that are generated from application (App) usage records provide more valuable information than mobility features since these emerging Apps heavily influence people’s lifestyles, such as their shopping, communication and education. Some previous studies found that there are substantial differences in the App usage patterns for different subscribers [20, 21] and discovered that Apps usage can reflect people’s behaviors and preferred lifestyles [22]. Another study differentiated mobile phone users according to their used Apps [23]. According to this information, we expect that the online indicators that are extracted from Apps usage data should be associated with SELs. In addition, another study [24] discovered the functions of city regions according to human mobility and points of interests (POI), which inspires us to investigate whether the distribution of the POIs in a region may be an important indicator of SELs.

In this paper, we propose a novel approach to estimate SELs of blocks in Shanghai using the online and offline features of an urban region. The online and offline features that are extracted from App usage records and the distribution of the POIs in the region are entered into Random Forests (RF) or support vector machines (SVM) in order to predict SELs of blocks in Shanghai. Our approach achieves better performance due to considering the online features compared with previous work that just considered offline features. We find that frequent mobility corresponds to developed economic status. In addition, the area with a large quantity of finance POIs tends to be in high socio-economic levels. Furthermore, we find a tight correlation between App usage patterns and economic levels. People in wealthy areas are more active on networks and spend more time on Apps in travel category. This method provides a more precise approach to classification prediction than previous methods using calling details data [8, 9], which achieved the 86.6% accuracy for 3 SELs and 72.7% accuracy for 4 SELs in the blocks of Shanghai. In addition, it addresses the problem of high manpower and time costs when NSI estimates the SELs of regions. Our research provides reliable and timely predictions for policy makers to allocate resources for regions with different SELs, which can promote the equality of economic growth and improve people’s livelihoods.

Materials and methods

Datasets

Shanghai is a metropolis in China that is known as a world-famous financial center. It is composed of 16 districts and divided into 9858 regions according to the coverage of the base
stations. However, the base station level is extremely fine-grained. In addition, we cannot obtain the GDP at the base station level. Thus, we investigate the SELs of the urban regions in Shanghai at a coarser granularity and aggregate 9858 base stations into 188 blocks according to their location information. The block is an administrative area lower than a district. The Shanghai government divides the 16 districts into 215 blocks for administrative ease. Our datasets cover 188 blocks in 15 districts except for the Chongming district. The boundary of each block consists of discrete points with longitudes and latitudes. Comparing the base station location data to the boundary location data of blocks, we assign the 9858 base station regions into 188 blocks. We use POIs data, App usage records and the GDPs of the urban regions in the SEL prediction. For each block in Shanghai, the features that are entered into prediction model contain over 6000 online and offline features, including App features, flow features, POI features and origin-destination (OD) features. In next section, we illustrate how to extract features for blocks from App usage records and POIs data.

**POIs data.** We obtained POIs in Shanghai which refer to locations that are associated with certain functions such as companies, restaurants, schools, government agencies, etc. We crawled the 782,528 POIs of Shanghai from Baidu Map service and complied with the terms and conditions of the platform. According to the functions of POIs, we divide the POIs into 21 categories. There are restaurant, hotel, shopping mall, entertainment, industry, fitness, office building, residence, view spot, travel, education, finance, company, factory, life service, technology park, economic development zone, high-tech development zone, town, business and village. The POIs data include 9,858 records. Each record consists of the base station ID and the number of each POI category within the coverage of the base station. The POI distribution reflects the offline characteristics of the regions.

**App usage records.** We obtained a historical log of anonymous App usage records in Shanghai from China Telecom, which is one of the major mobile operators in China. Our dataset contains anonymous cellular data access traces that were obtained using Deep Packet Inspection (DPI) appliances. DPI is a popular approach that is used for accurately detecting traffic in terms of the content, Apps, and individual users [25]. After DPI, each access log to the cellular network is characterized by an anonymous user ID, timestamp, cellular base station with GPS location and connection metadata. The ethical issues of the data have been considered carefully. All the App usage records that are used in our research are anonymized. In addition, all the datasets are stored in a server that is not connected to the Internet. Furthermore, all the researchers that participate in this work signed a Non-Disclosure Agreement and agree not to use the data for any other purpose. More details about how to obtain data from China Telecom and how to protect user privacy are illustrated in the S1 File.

The App usage records cover 15 districts of Shanghai, with the exception of the Chongming district. The data set contains the records of 20,000 subscribers over a period of one week from April 20th to April 26th, 2016. The logs contain 2,000 unique Apps and cover 9,858 base stations. Each log has the form of a tuple of the user ID, timestamp, base station ID, and App ID. The records with spatio-temporal information can be used to reveal human mobility patterns. With the locations of the base stations that a user went through, the user’s mobility can be generated as offline features. Moreover, the App usage records reflect users’ online characteristics [22, 26, 27]. We can observe the online features, including the category of the and the duration use.

**GDP of urban regions.** To validate the prediction model, the ground truth data are obtained from the *Shanghai Economy Almanac* (2017) [28] that was edited by the Development Research Center of Shanghai Municipal People’s Government and The Sixth National Census. The almanac is the most systematic, complete and authoritative reference book that records the economic and social development of Shanghai. From this book, we obtained the
GDPs of 15 administrative districts of Shanghai in 2016. Moreover, we obtained the populations of the 188 blocks from The Sixth National Census. To achieve a fine-grained prediction of the regions in Shanghai, we computed the GDPs of the blocks using the GDPs of the 15 administrative districts and populations of 188 blocks. The calculation formula is displayed as follows:

$$G_b = \frac{G_d P_b}{P_d}$$

where $G_b$ represents the GDP of a block and $G_d$ represents the GDP of the district to which the block belongs, and $P_b$ and $P_d$ are populations of district and block, respectively.

With the GDPs of the 188 blocks in Shanghai, we calculate the average GDP of the blocks and define 3 SELs ($A, B, C$) and 4 SELs ($A, B, C, D$), respectively, according to the difference between the GDPs of the blocks and the average GDP, where $A$ represents the highest socio-economic level. The GDPs of the 188 blocks range from 21.75 to 671.11 with an average of 142.66. For the 5th block lacks POIs data, we drop it when training the model. For 3 classes, $A$ covers the range [100 \sim 300), $B$ covers [-100 \sim 100) and $C$ covers [-300 \sim -100). For 4 classes, $A$ covers the range [100 \sim 300), $B$ covers [0 \sim 100), $C$ covers [-100 \sim 0) and $D$ covers [-300 \sim -100). There are a few blocks whose GDPs are higher than all the defined SELs. We just assign them to the highest level. The distribution of blocks across 3 classes is 26 blocks in level $A$, 148 blocks in level $B$ and 13 blocks in level $C$. The distribution of blocks across 4 classes is 26 blocks in level $A$, 39 blocks in level $B$, 109 blocks in level $C$ and 13 blocks in level $D$.

Feature engineering

The objective of feature engineering is to generate indicators that can be used to predict the SELs of regions. In the existing work, the mobility features are regarded as effective indicators for predicting socio-economic status. In our work, we seek to explore more effective indicators of SELs. Apart from mobility features [24, 29, 30] such as the flow features and OD features, we extract the POI features and App features from row POIs data and App usage records. The offline features consist of POI features, flow features and OD features, while the online features represent the App features. In this paper, we extract the features that are mentioned above from App usage records and the POIs data at the block level and enter them into the RF or SVM to predict SELs of the blocks in Shanghai.

Baseline models: Mobility features, offline features and random classifier. In addition to the model we proposed below, we introduce three baselines for comparison purposes. The first baseline is a mobility based model, which consists of the 5 most important flow features and the 5 most important OD features. The second baseline is a model using offline features, which consists of the mobility features that were used above and the 10 most important POI features. The third baseline is a random classifier whose features are the same as our model’s features. The difference between random classifier and our model is that the training sets of random classifier are chose randomly. Since the distribution of blocks at different SELs is unbalanced, we use cost-sensitive to solve this problem in our model. Meanwhile, in the random classifier, it chooses the training samples from each class randomly. Next, we illustrate how to extract these features and evaluate the importance of features using formula (2).

Jing Yuan et al. found that the functions of different city regions are correlated with the POIs in them. We expect to find that the distribution of the POIs within blocks is effective for predicting SELs. We then aggregate the POI distribution of the base stations at the block level by comparing the locations of the base stations with the boundaries of each block. Since the 5th block lacks POI data, we exclude it and assigned the 9858 base station records to 187
blocks. Ultimately, for each block, the POI feature is a vector consisting of number of POIs belonging to the 21 POI categories.

The flow features represent the incoming and outgoing visitor flowrates of the regions. We assign the 9,858 base stations to 188 blocks and observe the flowrate of these blocks during 8 three-hour periods. To avoid the situation that the subscribers just pass through a location but do not stay for long, we set a time threshold $\Delta T$ to define whether they stay. In our research, we set $\Delta T = 30$ minutes according to the experience of previous work [31]. Only when the user stayed in a block for a time greater than $\Delta T$ can the block be treated as an effective stay point. Thus, the flow feature vector of each block is represented by a $2 \times 2 \times 8$ (time periods) vector whose element represents the average daily flowrate at a certain period. For instance, once an individual leaves stay point $j$ at AM 8:00 and arrives at stay point $i$ at AM 10:00 on Sunday, a value of 1 will be added at $OUT_H06 - 09$ of block $j$ and $IN_H09 - 12$ of block $i$, respectively.

We observed transitions between the original block and destination block as the OD features. We defined the OD features as transitions where people go to or leave a block during a certain period [9]. For instance, given two App usage records of the same subscriber, an OD feature can be denoted as follows. If a subscriber leaves stay point $j$ at AM 8:00 and arrives at stay point $i$ at AM 10:00 on Sunday, a value of 1 will be added at $OUT_i_H06 - 09$ of block $j$ and $IN_j_H09 - 12$ of block $i$, respectively. Ultimately, the OD feature of a block is represented by a $2 \times 2 \times 8 \times 188$ (origination or destination) average mobility motif tuple.

Our model. Our model combines all the offline features that are mentioned above and the App features. We then input them into the machine learning algorithms in order to train the prediction model. Apps that people pay attention to vary according to their role(s). For example, student groups may spend more time on education Apps, while white-collars may pay more attention to financial Apps. Thus, we assume that the App usage preferences of the people in a region may have some correlations with the SELs of the region. The App features are extracted from App usage records as follows.

There are 2,000 unique Apps in the App usage records. The number of Apps is extremely large and some of the Apps have similar functions. Thus, the Apps are divided into 18 categories, as shown in Table 1, according to their core functions, which are provided by the App store. The motivation for why we extract App features is that App usage can reflect users’ behaviors and preference. However, the two categories of System_tool and Others cannot provide valuable feedback. Therefore, we drop these two categories when we extract the App features. In order to further observe the influence of different time periods on App usage, we separate a week into weekdays and holidays. Each day is divided into 8 three-hour periods i.e.,

| App Category | Sample Apps    | App Category | Sample Apps     |
|--------------|----------------|--------------|-----------------|
| Social       | QQ, Wechat     | Finance      | Alipay, Ant fortune |
| Video        | QIYIVideo, Pear Video | Transportation | Uber, Didi     |
| Audio        | QQMusic, QingtingFM | Navigation | Amap, Baidumap |
| News         | SohuNews, QQNews | Life         | Clouds weather |
| Browser      | UC Browser     | Travel       | Qunar, Fliggy  |
| Game         | Happyelement   | Office       | Iciba, QQMail  |
| Health       | Keep, Yuedongli | Stock        | Eastmoney      |
| Shopping     | Taobao, JD     | System tool  | WiFi, Sogoulput |
| Ordering     | Meituan, Eleme | Others       | Androidasync   |

https://doi.org/10.1371/journal.pone.0219058.t001
\( T \in \{[0 \sim 3], [3 \sim 6], \ldots, [21 \sim 24]\} \). For each block, the App feature is represented by a 2 (weekdays and holidays)×8 (time periods)×16 (App categories) vector. For example, the element in \( \text{financial}_H12 \sim 15 \) of block \( i \) reflects the average usage duration of the \text{financial} Apps in the 12:00-15:00 PM period on holidays in block \( i \).

**Results**

**Accuracy of prediction**

**Accuracy of prediction model.** The GDPs of the 188 blocks ranged from 21.75 to 671.11 with an average of 142.66. For the 5th block lacks POIs data, we drop it when training the model. We divide the 187 blocks into 5 different training and testing sets through 5-fold cross-validation. We fit a regression algorithm using the Random Forest to each set and obtain the average \( R^2 \) of 0.474 as shown in Fig 1, which indicates that there is correlation between features that are extracted above and the socio-economic indicators. Since the previous works usually solve the economic prediction as classification problems, we will further analyze the results of the classification models.

Figs 2 and 3 show the socio-economic maps when classifying 187 blocks into 3 SELs and 4 SELs, respectively, in Shanghai. Fig 2(a) and 2(b) are the distributions of the actual SELs and the predicted SELs, respectively, with an accuracy of 85% for the 3 classes. Fig 3(a) and 3(b) are respectively the distributions of the actual SELs and the predicted SELs with an accuracy of 64% for the 4 classes. The darker shades represent richer areas. The blocks without a filling color lack data. From the actual SEL map, we can observe that classification of SELs is unbalanced. Thus, we apply cost-sensitive training when we train the models. The predicted SEL maps are obtained using SVM model through a 5-fold cross-validation. The model has good performance though the classification is unbalanced.

![Fig 1. Comparison of GDP predictions to actual GDP.](https://doi.org/10.1371/journal.pone.0219058.g001)
Comparison with baselines. We use 5-fold cross-validation to divide the dataset into 5 training and testing sets with an 80% – 20% split. We fit a RF or SVM on each training set and observe the model’s performance on the testing sets. Tables 2 and 3 show the accuracy (ACC), precision (P), recall (R) and F1 scores of different models for the 3 SELs and 4 SELs, respectively. For both the 3 classes and 4 classes, the trends of the performance indicators are similar. Compared with model using mobility features, the offline feature based model has approximately 4% better accuracy for the 3 SELs and approximately 3% better accuracy for the 4 SELs when using RF. Furthermore, the random classifiers evidently enhance the overall performance both with 3 SELs and 4SELs. However, the random classifier for the 3 SELs performs badly when predicting the blocks in SEL C, which is influenced by the unbalanced classes. Our models make up for this limitation to a certain extent and perform best on all aspects. In the case of our method, the highest accuracies for the 3 SELs and 4 SELs are obtained by using the SVM, while the RF models work better on detecting the blocks of SEL C or SEL D. The results indicate that POI features and App features play important parts in enhancing predictive performance.

Fig 2. The actual and predicted socio-economic map of 3 SELs. a: Actual 3 SEL map. b: Predicted 3 SEL map with the accuracy of 85%.
https://doi.org/10.1371/journal.pone.0219058.g002

Fig 3. The actual and predicted socio-economic map of 4 SELs. a: Actual 4 SEL map. b: Predicted 4 SEL map with the accuracy of 64%.
https://doi.org/10.1371/journal.pone.0219058.g003
To further investigate how the model performance is affected by the number of features that is entered into the model, we compute the accuracy using 5-fold cross-validation for each subset of ordered features using RF and SVM. The ordered features are selected according to the ANOVA F-value of the samples. As shown in Fig 4(a), we observe that the RF for the 3 SELs achieves an accuracy of 84.5% when using the top 31 features, while the RF for the 4 SELs reaches 65.7% when using the top 211 features.

### Table 2. Accuracy, precision, recall and F1-score of different models for 3 SELs.

| SEL  | Mobility features | Offline features | Random classifier | Our model |
|------|------------------|------------------|------------------|-----------|
|      | P/R/F1           | P/R/F1           | P/R/F1           | ACC/P/R/F1|
| SEL A | RF 0.57/0.31/0.40 | RF 0.67/0.38/0.49 | RF 0.80/0.46/0.59 | RF 0.91/0.38/0.54 |
| SEL B | SVM 0.44/0.54/0.48 | SVM 0.83/0.97/0.89 | SVM 0.84/0.98/0.91 | SVM 0.86/0.99/0.92 |
| SEL C | RF 0.30/0.54/0.39 | SVM 0.85/0.55/0.67 | SVM 0.82/0.89/0.85 | RF 0.86/0.99/0.92 |
| Total | SVM 0.78/0.73/0.78 | SVM 0.75/0.81/0.76 | SVM 0.75/0.81/0.76 | SVM 0.86/0.85/0.87 |

### Table 3. Accuracy, precision, recall and F1-score of different models for 4 SELs.

| SEL  | Mobility features | Offline features | Random classifier | Our model |
|------|------------------|------------------|------------------|-----------|
|      | P/R/F1           | P/R/F1           | P/R/F1           | ACC/P/R/F1|
| SEL A | RF 0.43/0.35/0.38 | RF 0.59/0.38/0.47 | RF 0.89/0.31/0.46 | RF 0.81/0.50/0.62 |
| SEL B | SVM 0.48/0.54/0.51 | SVM 0.18/0.08/0.11 | SVM 0.38/0.46/0.41 | SVM 0.89/0.31/0.46 |
| SEL C | RF 0.61/0.81/0.70 | RF 0.63/0.86/0.73 | RF 0.67/0.89/0.76 | RF 0.68/0.92/0.78 |
| SEL D | SVM 0.25/0.08/0.12 | SVM 0.00/0.00/0.00 | SVM 0.25/0.08/0.12 | SVM 0.33/0.08/0.12 |
| Total | RF 0.78/0.73/0.74 | RF 0.72/0.46/0.56 | RF 0.67/0.83/0.74 | RF 0.68/0.96/0.80 |
|       | SVM 0.54/0.47/0.49 | SVM 0.23/0.85/0.36 | SVM 0.21/0.46/0.29 | SVM 0.20/0.08/0.11 |
|       | SVM 0.54/0.47/0.49 | SVM 0.45/0.56/0.45 | SVM 0.58/0.61/0.59 | SVM 0.72/0.68/0.63 |

Accuracy versus the number of features. To further investigate how the model performance is affected by the number of features that is entered into the model, we compute the accuracy using 5-fold cross-validation for each subset of ordered features using RF and SVM. The ordered features are selected according to the ANOVA F-value of the samples. As shown in Fig 4(a), we observe that the RF for the 3 SELs achieves an accuracy of 84.5% when using the top 31 features, while the RF for the 4 SELs reaches 65.7% when using the top 211 features.

![Accuracy of model with RF](https://doi.org/10.1371/journal.pone.0219058.g004)  
![Accuracy of model with SVM](https://doi.org/10.1371/journal.pone.0219058.g004)
The classification results using the SVM are presented in Fig 4(b). The SVM model for the 3 SELs achieves the best result, reaching an accuracy of up to 86.6% by using the top 381 features, while the model for the 4 SELs achieves the highest accuracy of 72.7% for the top 1081 features. Thus, the SVM outperforms the RF both for 3 classes and 4 classes.

Comparisons between rich and poor areas

We compute the importances of all the features when using the RF model and obtain 1018 features whose scores are greater than 0. The importance of feature $x$ is defined as follows:

$$I(x) = E_s(g) - E^{o}(g)$$

where $E_s(g)$ is the out of bag (OOB) error of the RF model, while the $E^{o}(g)$ is the new OOB error when the values of feature $x$ are inserted noise. The higher importance value of the feature is, the more important is the feature. The distribution of the feature scores is presented in Fig 5, which shows that the features with high scores just account for a small proportion. Table 4 shows the top 20 important features according to their categories and their scores. As shown in Table 4, the online features, flow features and POI features of offline features have significant importance, and we simply compare these features between rich areas and poor areas.

To further explore the correlation between the features and SELs, we selected the 5 richest and 5 poorest blocks in Shanghai. The rich blocks are administrative centers or economic parks, while the poor blocks mainly rely on agriculture. For convenience and to provide a clear illustration, we compute the average of the selected features of rich blocks and the poor blocks, respectively, and make comparisons.

The comparison of offline features. As revealed in Table 4, there are 3 POI features in the top 10 features, which indicates that POI features have good correlations with SELs. For each block, the POI feature is a vector consisting of the number of 21 categories of POIs.
We select the 5 most important POI features and make a comparison between rich areas and poor areas. The differences in the POI features between rich areas and poor areas are shown as Fig 6(a).

Apparently, the **finance** POIs have the closest correlation with economics. The gap between rich areas and poor areas with respect to **finance** POIs demonstrates that there are more financial business requirements for people in rich areas, from which we can infer that regions with more financial POIs may have higher socio-economic level. Moreover, more **education** and **office building** POIs are located in rich areas compared with poor areas. Therefore, more opportunities for learning and work are provided in richer areas. In addition, the difference between rich areas and poor areas with respect to **hotel** POIs is apparent, which indicates that the people visiting a city are more likely to stay in the richer area of the city. What is contrary to our expectations is that the rich areas have more villages. That may be because Shanghai is a highly urbanized city whose villages are not dominated by traditional agriculture. Thus, the large number of villages corresponds to the population to some extent.

### Table 4. Top 20 important features for the 3 classes using the RF

| Feature                | Score  | Feature                | Score  |
|------------------------|--------|------------------------|--------|
| IN_188_W15-18          | 0.011  | Transportation_W09-12  | 0.0063 |
| Navigation_W21-24      | 0.0096 | Social_H21-24          | 0.0061 |
| Finance POIs           | 0.009  | Browser_H15-18         | 0.006  |
| Browser_W21-24         | 0.0083 | IN_H18-21               | 0.0059 |
| Navigation_W06-09      | 0.0081 | IN_H21-24               | 0.0056 |
| Education POIs         | 0.0075 | IN_W15-18               | 0.0055 |
| Social_W18-21          | 0.0072 | News_W12-15             | 0.0052 |
| Office building POIs   | 0.0071 | Browser_H12-15          | 0.0051 |
| IN_W09-12              | 0.0066 | OUT_W12-15              | 0.0051 |
| OUT_H18-21             | 0.0064 | IN_W21-24               | 0.0051 |

We select the 5 most important POI features and make a comparison between rich areas and poor areas. The differences in the POI features between rich areas and poor areas are shown as Fig 6(a).

Apparently, the **finance** POIs have the closest correlation with economics. The gap between rich areas and poor areas with respect to **finance** POIs demonstrates that there are more financial business requirements for people in rich areas, from which we can infer that regions with more financial POIs may have higher socio-economic level. Moreover, more **education** and **office building** POIs are located in rich areas compared with poor areas. Therefore, more opportunities for learning and work are provided in richer areas. In addition, the difference between rich areas and poor areas with respect to **hotel** POIs is apparent, which indicates that the people visiting a city are more likely to stay in the richer area of the city. What is contrary to our expectations is that the rich areas have more villages. That may be because Shanghai is a highly urbanized city whose villages are not dominated by traditional agriculture. Thus, the large number of villages corresponds to the population to some extent.
Areas with larger populations are more likely to be well developed. Apparently, the significant differences in the numbers of POIs between rich areas and poor areas correspond to the differences in the SELs. Thus, the categories of POIs are valuable indicators to socioeconomic levels.

For each block we investigated, the flow feature is a vector whose element represents the average incoming or outgoing visitor flowrate during a given period. Fig 6(b) shows the differences between rich areas and poor areas for the 5 most important flow features. From this comparison, we can observe that more people come in or go out on holiday evenings in rich areas, which indicates that people in rich areas actively engage in activities in the evening. For instance, people are likely to go to rich area for shopping, entertainment or leisure on holidays. In addition, on workdays, more people flow out during the 12:00-15:00 PM period and flow in during the 15:00-18:00 PM period. This phenomenon reflects the fact that people in rich areas have more career opportunities. In brief, the inflows and outflows of a region during a given period can accurately reflect the social economic levels to a great extent. Some other researchers also have found that people’s mobility is tightly linked with their economic status [32, 33].

The comparison of online features. Online features represent App features. Therefore, we select the top 5 most important App features and make a comparison between rich areas and poor areas. For each block, the App feature is a vector in which element is the average usage duration of a category of Apps during a given period in the block. The results in Fig 7 show the distinction between rich and poor areas.

As observed in Fig 7, people in rich areas primarily use navigation and transportation Apps, especially in the mornings and evenings on workdays. That could be caused by the fact that there are more working persons using navigation and transportation Apps on their way to work in rich areas. Accordingly, areas with larger workforces correspond to higher GDPs. In addition, on holiday evenings, people in rich areas spend more time on social and browser Apps, which indicates that they are more active on the Internet. The significant differences between rich areas and poor areas shows that these features are highly indicative of the SELs. To further investigate the correlation between App usage and SEL, we select the 5 most frequently used App categories in rich areas and poor areas, respectively, as shown in Fig 8.

![Fig 7. The comparison of App features between rich areas and poor areas.](https://doi.org/10.1371/journal.pone.0219058.g007)
Comparing Fig 8(a) with 8(b), we can find that the App usage categories are more diverse for people in rich areas. People in rich areas spend more time on travel and navigation Apps, from which we can infer that people in rich areas may engage in more public activities and they pay more attention to the quality of life. Conversely, people in poor areas primarily use music and news Apps. Their hobbies are relatively monotonous compared with people in rich areas. Furthermore, we explore the influence brought by App usage without considering the different periods, as shown in Fig 8(c) and 8(d). Comparing Fig 8(c) with 8(d), we find that people in rich areas pay more frequent attention to networking Apps. Simultaneously, they spend more time on Apps that are required when taking trips. Thus, Apps related to taking trips are strong indicators to economic status. Fig 8(d) shows that people in poor areas get their entertainment usually through music and video Apps. Social and browser Apps are also popular with people in poor areas, but the time they spent on these types of Apps is truly shorter than that of people in rich areas.

**Discussion and conclusion**

The model we proposed performs better than the three simple baselines. The POI features and App features are showed to be tightly linked to economic status. We select some features according to their importance and analyze them. We find that POIs are good indicator to SELs. Areas with more finance, education and office building POIs are more likely to be of good economic status. Understandably, areas with more numerous opportunities and functions are likely to achieve high GDPs. Moreover, the App features reveal the fact that regions in which people frequently use Apps about taking trips during rush hour are more economically developed. Corresponding to the App features, areas with more people flowing in and out usually have high socio-economic levels, especially during rush hour and weekend evenings.
This result reaffirms the rule that the areas with denser flows correspond to high socio-economic levels. Although our research provides a novel method to investigate the socio-economic levels, there are still some limits. It is difficult to collect App usage records in the areas where smartphones are not ubiquitous. In addition, our method has an inherent limitation in that it does not capture smartphone Apps that are used through WiFi networks. Thus, the online features are not compete to a certain extent. Our method may provide some inspirations for researchers who are investigating the Internet data and socio-economic problems.

Supporting information

S1 File. Data description and construction. Supplementary text for this article. (PDF)

Acknowledgments

The authors thank the China Telecom for access to the data.

Author Contributions

Conceptualization: Xiang Chen.
Data curation: Yi Ren.
Formal analysis: Yi Ren, Tong Xia, Yong Li.
Funding acquisition: Xiang Chen.
Investigation: Yi Ren, Yong Li, Xiang Chen.
Methodology: Yi Ren, Tong Xia, Yong Li.
Project administration: Yong Li, Xiang Chen.
Software: Yi Ren.
Supervision: Yong Li, Xiang Chen.
Validation: Yi Ren.
Visualization: Yi Ren, Tong Xia.
Writing – original draft: Yi Ren.
Writing – review & editing: Tong Xia, Yong Li, Xiang Chen.

References

1. Luo S, Morone F, Sarraute C, Travizano M, Makse HA. Inferring personal economic status from social network location. Nature Communications. 2017; 8:15227. https://doi.org/10.1038/ncomms15227 PMID: 28509896
2. Blumenstock J, Cadamuro G, On R. Predicting poverty and wealth from mobile phone metadata. Science. 2015; 350(6264):1073. https://doi.org/10.1126/science.aac4420 PMID: 26612950
3. Smith-Clarke C, Mashhadi A, Capra L. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks; 2014.
4. Ebener S, Murray C, Tandon A, Elvidge CC. From wealth to health: modelling the distribution of income per capita at the sub-national level using night-time light imagery. International Journal of Health Geographics. 2005; 4(1):5. https://doi.org/10.1186/1476-072X-4-5 PMID: 15705196
5. Einav L, Levin J. Economics in the age of big data. Science. 2014; 346(6210):1243089. https://doi.org/10.1126/science.1243089 PMID: 25378629
6. Gutierrez T, Krings G, Blondel VD. Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. Computer Science. 2013;

7. Frias-Martinez V, Soto V, Virseda J, Frias-Martinez E. Computing cost-effective census maps from cell phone traces. In: Workshop on pervasive urban applications; 2012.

8. Soto V, Frias-Martinez V, Virseda J, Frias-Martinez E. Prediction of Socioeconomic Levels Using Cell Phone Records. Lecture Notes in Computer Science. 2011; 6767(2):377–388. https://doi.org/10.1007/978-3-642-22362-4_35

9. Hong L, Frias-Martinez E, Frias-Martinez V. Topic models to infer socio-economic maps. In: Thirtieth Aaai Conference on Artificial Intelligence; 2016.

10. Pappalardo L, Pedreschi D, Smoreda Z, Giannotti F. Using Big Data to study the link between human mobility and socio-economic development. In: IEEE International Conference on Big Data; 2015.

11. Elvidge CD, Baugh KE, Kihn EA, Kroehl HW, Davis ER. Mapping city lights with nighttime data from the DMSP Operational Linescan System. Eng & Remote Sens. 1997; 63(6):727–734.

12. Jean N, Burke M, Xie M, Davis WM, Ermon S. Combining satellite imagery and machine learning to predict poverty. Science. 2016; 353(6301):790. https://doi.org/10.1126/science.aaf7894

13. Nathan E, Michael M, Rob C. Network diversity and economic development. Science. 2010; 328(5981):1029–1031. https://doi.org/10.1126/science.1186605

14. Varian HR. Big Data: New Tricks for Econometrics. Journal of Economic Perspectives. 2014; 28(2):3–27. https://doi.org/10.1257/jep.28.2.3

15. Varian HR, Choi H. Predicting the Present with Google Trends. Economic Record. 2012; 88(s1):pags. 2–9.

16. Choi H. Predicting Initial Claims for Unemployment Benefits. Social Science Electronic Publishing. 2010;

17. Granovetter M. The Impact of Social Structure on Economic Outcomes. Journal of Economic Perspectives. 2005; 19(1):33–50. https://doi.org/10.1257/0895330053147958

18. Pappalardo L, Vanhoof M, Gabrielli L, Smoreda Z, Pedreschi D, Giannotti F. An analytical framework to nowcast well-being using mobile phone data. International Journal of Data Science & Analytics.2016; 2(1-2):75–92. https://doi.org/10.1007/s41060-016-0013-2

19. Pan W, Ghoshal G, Krumme C, Cebrian M, Pentland A. Urban characteristics attributable to density-driven tie formation. Nature Communications. 2012; 4(3):1961.

20. Xu Q, Erman J, Gerber A, Mao Z, Pang J, Venkataraman S. Identifying diverse usage behaviors of smartphone apps. In: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. ACM; 2011. p. 329–344.

21. Falaki H, Mahajan R, Kandula S, Lymberopoulos D, Govindan R, Estrin D. Diversity in smartphone usage. In: Proceedings of the 8th international conference on Mobile systems, applications, and services. ACM; 2010. p. 179–194.

22. Zhao S, Ramos J, Tao J, Jiang Z, Li S, Wu Z, etÂ. al. Discovering different kinds of smartphone users through their application usage behaviors. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM; 2016. p. 498–509.

23. Welke P, Andone I, Blaszkiewicz K, Markowitz A. Differentiating smartphone users by app usage. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM; 2016. p. 519–523.

24. Yuan J, Zheng Y, Xie X. Discovering regions of different functions in a city using human mobility and POIs. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2012. p. 186–194.

25. Kumar S, Dharmapurikar S, Fang Y, Crowley P, Turner J. Algorithms to accelerate multiple regular expressions matching for deep packet inspection. Acm Sigcomm Computer Communication Review. 2006; 36(4):339–350. https://doi.org/10.1145/1151659.1159952

26. Li H, Xuan L, Liu X, Tao X, Bian K, Lin FX, et al. Characterizing Smartphone Usage Patterns from Millions of Android Users. In: Internet Measurement Conference; 2015.

27. Canneyt SV, Bron M, Haines A, Lalmaz M. Describing Patterns and Disruptions in Large Scale Mobile App Usage Data. In: the 26th International Conference; 2017.

28. Shanghai Economy Almanac; Development Research Center of Shanghai Municipal People’s Government; 2017.

29. Zhang K, Lin YR, Pelechrinis K. EigenTransitions with Hypothesis Testing: The Anatomy of Urban Mobility. In: Tenth International Aaai Conference on Web & Social Media; 2016.
30. Pappalardo L, Vanhoof M, Gabrielli L, Smoreda Z, Pedreschi D, Giannotti F. An analytical framework to nowcast well-being using mobile phone data. International Journal of Data Science & Analytics. 2016; 2 (1-2):75–92. https://doi.org/10.1007/s41060-016-0013-2

31. Zheng Y, Xie X, Ma WY. Geolife: A collaborative social networking service among user, location and trajectory. IEEE Data Eng Bull. 2010; 33(2):32–39.

32. Rubio A, Frias-Martinez V, Frias-Martinez E, Oliver N. Human Mobility in Advanced and Developing Economies: A Comparative Analysis. 2010;.

33. Friasmartinez V, Virseda J, Friasmartinez E. 1 Socio-Economic Levels and Human Mobility. In: Qual Meets Quant Workshop-qmq at the Int Confon Information & Communication Technologies & Development; 2013.