Predicting Socio-Economic Levels of Individuals via App Usage Records

Yi Ren[1,2], Weimin Mai[1,2], Yong Li[3], and Xiang Chen[1,2(✉)]

[1] School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China
chenxiang@mail.sysu.edu.cn
[2] Key Lab of EDA, Research Institute of Tsinghua University in Shenzhen (RITS), Shenzhen 518075, China
[3] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

Abstract. The socio-economic level of an individual is an indicator of the education, purchasing power and housing. Accurate and proper prediction of the individuals is of great significance for market campaign. However, the previous approaches estimating the socio-economic status of an individual mainly rely on census data which demands a great quantity of money and manpower. In this paper, we analyse two datasets: App usage records and occupation data of individuals in a metropolis of China. We divide the individuals into 4 socio-economic levels according to their occupations. Then, we propose a low-cost socio-economic level classification model constructed with machine learning algorithm. Our predictive model achieves a high accuracy over 80%. Our results show that the features extracted from user's App usage records are valuable indicators to predict the socio-economics levels of individuals.

Keywords: Data mining · Mobile data · Socio-economics level

1 Introduction

The socio-economic level (SEL) is an important indicator used to characterize the social status of the individuals in sociology. With the information about people's social status, the company can estimate the consumer's purchasing power and implement precision marketing to different consumer groups. In addition, from a public perspective, socio-economics status of individuals is useful for making social proper policies. Some studies found that people with different socio-economics levels might have various scenarios in choosing health services [22] and the ways of transportation [13].

The current approaches of investigating the SELs of individuals usually depend on the economics survey constructed by National Statistical Institute (NSI). The data obtained from the survey is detailed and authoritative. However, this method has some limitations. For the government leading the economics survey, it would demand much money and manpower to obtain the data

X. B. Zhai et al. (Eds.): MLICOM 2019, LNICST 294, pp. 199–210, 2019.
https://doi.org/10.1007/978-3-030-32388-2_17

of all the residents. In addition, NSI holds the economics survey every 5 years in China, thus not being able to provide timely changes about human's economics status. Furthermore, for the areas in poor economic status, it might have no access to obtain the economics data. Therefore, a low-cost and timely method to estimate the SELs of individuals is needed.

Many researchers have been attempting to investigate the individual's SEL using novel data and methods [14,17,19]. Previous work has found that social network influents economic status [7]. Shaojun Luo et al. extracted social networks of users from their telecommunications [14]. Then, they inferred personal economic status from their social networks. However, with the ubiquitous usage of smart phones and applications (App), people gradually connect with others using chat Apps instead of callings. Therefore, the calling data used to generate users' social networks is being outdated. Some other studies [9,10,21] predicted SELs of individuals by using data collected from users' social media, such as Twitter and Sina Weibo. However, there exist some users just browsing Blogs instead of updating their own microblog. Therefore, it's difficult to characterise the users who are not active in social media.

In this paper, we analyse two datasets: the App usage records of users in Shanghai and their occupation information collected from Sina Weibo. We divide the users into 4 SELs based on their occupations. We extract users' App features and mobility features from their App usage records. Then we enter these features into machine learning algorithms to train the personal SEL predictive model. Our model achieves an accuracy over 80% when classifying the SELs of users.

2 Datasets

In this research, we investigate and predict the SELs of some individuals in Shanghai. Firstly, we obtain the App usage records of 1,200 users from China Telecom. Secondly, we crawl the Sina Weibo identifications of these users according to the user ID that corresponding to their App usage records. Then, we obtain 660 effective users whose Sina Weibo identifications contain their occupation information. At last, we divide the 660 users into 4 socio-economic levels and train a classification predictive model.

Data privacy has been protected through many measures. Firstly, we use reconstructed user ID instead of the actual identification in order to protect the privacy of the users. Secondly, all the datasets used in this research are stored in off-line server. People who are not related to this study were unable to access these datasets. Thirdly, all the researchers signed Non-disclosure agreement.

2.1 App Usage Records

The App usage records used in this research contain 1,200 users' information during a week from April 20th to April 26th, 2016. The dataset contains 2,000 applications and covers 9,858 base stations distributed in 15 districts of Shanghai. Each record consists of user ID, timestramp, base station ID and App ID

as shown in Table 1. With the time and geographic information of users, we can discover their mobility behaviors. In addition, the records can reflect human's preferences and habits in using Apps.

Table 1. The examples of app usage records.

User ID	Timestramp	Base station ID	App ID
1001628504	20160421171508	3610000B2B02	2
1001628504	20160421171518	3610000B2B02	2
1001628504	20160421171535	3610000B2B03	2
1002143827	20160419192243	3610000C21A3	5
1002143827	20160419213052	3610000C21A3	5
1002143827	20160419213430	3610000C21A3	1120

2.2 Occupation Information

In order to validate the performance of our classification model, we crawl the occupation contained in the Sina Weibo identification of each user as ground truth. At last, 660 users with occupation are obtained and regarded as effective data. Usually, the sociology researchers stratify people into 5 levels according to their occupations [11]: upper socio-economic level, upper middle socio-economic level, middle socio-economics level, lower middle socio-economic level, lower socio-economic level. Qiang Li et al. surveyed 99 occupations in Beijing and computed their corresponding scores [23]. They also divided 99 occupations into 5 classes. A subset of the social stratification is shown as Table 2. Beijing is a metropolis which is similar to Shanghai in terms of social development, population and urbanization level. Therefore, we used the stratified standard in Qiang Li's research to define the SELs of the 660 users.

Table 2. Samples of socio-economic stratification.

Socio-economic stratification	Occupation examples
Upper SEL	Scientist, bank president, lawyer
Upper middle SEL	Writer, translator, police
Middle SEL	Nurse, purchasing agent
Lower middle SEL	Hairdresser, insurance salesman
Lower SEL	Rickshaw puller, junkman

The 660 users in our study cover 4 SELs except the lower SEL. This may be caused by the fact that people with lower SEL have no need to apply for

Sina Weibo identification, or they don't use Sina Weibo. Thus, we divide 660 user into 4 SELs (SEL A, SEL B, SEL C, SEL D), where SEL A represents the highest level, SEL D represents the lowest level. The 660 users are distributed as follows, 45 users in SEL A, 497 users in SEL B, 112 users in SEL C and 7 users in SEL D.

3 Feature Engineering

In this section, we overview and generate the features of 660 users used in the SEL classification model. The features contain 2 categories, including App features and mobility features.

3.1 App Features

Previous researchers [12,26,27] found that people's App usage behaviors could reflect their attributes and traits. Therefore, we expect that people's App usage behaviors have correlations with their economic status. We extract people's App features from App usage records as the following steps. (1) We classify 2,000 Apps into 18 categories based on their core functions as shown in Table 3. For the categories of $SYSTEM\ TOOL$ and $OTHERS$ can not reflect users' preference, we drop these 2 categories when extracting features. (2) Previous App usage work [25] has found that the App usage patterns have differences in different time periods. Thus, we divide a week into workdays and weekends. Then, we divide a day into four 6-hour ranges i.e. $T \in \{[0-6), [6-12), [12-18), [18-24)\}$. (3) For each user, the App feature is a vector of 2(weekdays and weekends) $\times$ 4(time periods) $\times$ 16(App categories). Each element of the vector represents the time that the user spends on every App category during a certain period.

Table 3. App categories

App Category	Apps	App Category	Apps
SOCIAL	QQ	BROWSER	UC Browser
FINANCE	TongHuaShun	TRAVEL	Qvnar
VIDEO	iQiYi	GAME	Xiaoxiaole
TRANSPORTATION	Didi Taxi	OFFICE	163Mail
AUDIO	QQ Music	HEALTH	Keep
NAVIGATION	GaoDe Map	STOCK	Eastmoney
NEWS	QQNews	SYSTEM TOOL	Wifi
LIFE	Clouds Weather	OTHERS	Androdasync

3.2 Mobility Features

Luca Pappalardo et al. [18] found that there exist links between human mobility and socio-economic development. With the App usage records with time and space information, we can extract the mobility features of users. The distribution of 9858 base stations in Shanghai is shown with red points in Fig. 1. According to the longitudes and latitudes, we aggregate all the base stations into 188 blocks as shown in Fig. 2.

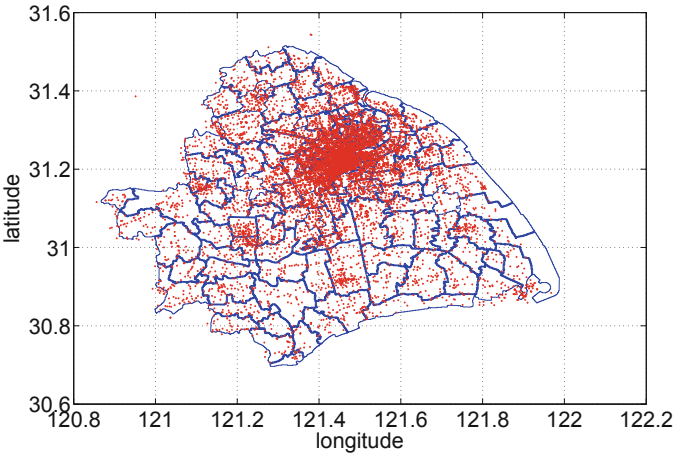

Fig. 1. The distribution of 9,858 base stations in Shanghai.

Table 4 introduces the 3 types of mobility features. m_1 represents the times of a user visiting 188 blocks in a week, respectively. The 188 blocks are in 3 socio-economic levels. m_2 represents the times of a user visiting rich, intermediate and poor blocks, respectively. In addition, we divide a day into 4 time periods as mentioned above. Then, we count the number of times a user visit blocks in 3 SELs during a certain period. m_3 is a tuple of 2(weekdays and weekends) $\times$ 3(3 SELs of blocks) $\times$ 4(time periods).

Table 4. Three types of mobility features.

m_1	Number of times a user visit 188 blocks in a week
m_2	Number of times a user visit rich, intermediate and poor blocks
m_3	Number of times a user visit blocks in 3 SEL during a certain period

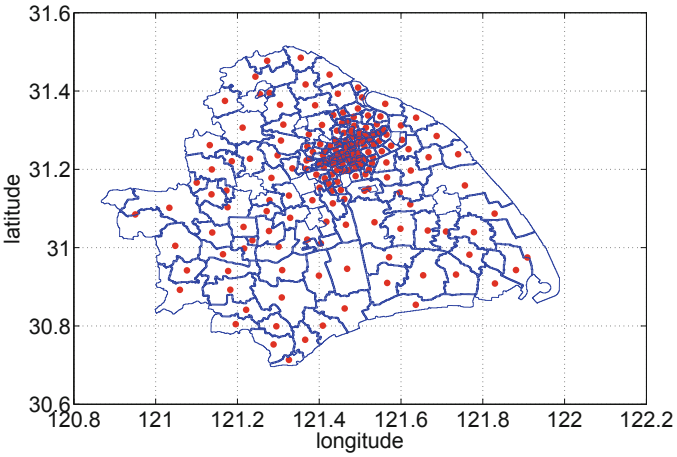

Fig. 2. The distribution of administrative blocks in Shanghai.

4 Classification with FM

In this section, we use the classification model based on Factorization Machine (FM) to train the predictive model. FM is an algorithm of solving binary classification predictive problem [24]. Comparing to the traditional linear model, FM concerns the impact brought by the interaction between the i-th and j-th features. FM model is shown as the following.

$$\hat{y}(x) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{ij} x_i x_j \tag{1}$$

where w_0 is the global bias, w_i represents the weight of the feature x_i, w_{ij} is the interaction between feature x_i and x_j. For the convenience of estimating the w_{ij}, we set $w_{ij} := \boldsymbol{v}\boldsymbol{v}^T$, where $\boldsymbol{v}_i = (v_{i1}, v_{i2}, ..., v_{ik})^T, i = 1, 2, ..., n$.

In order to discover insight, we train a multi-class classification model based on FM instead of using multi-classification algorithms. For each class, we train a one-vs-all binary classification with FM. We apply Stochastic Gradient Descent (SGD) to learn the parameters of FM model. Finally, we select the level which has the highest probability to be the SEL of the user.

5 Results and Discussions

Class-imbalance is one of the most common but difficult problems in data mining. A large number of researchers have been studying how to train proper model from class-imbalance datasets. As revealed in Sect. 2, the samples distributed in 4 SELs are extremely unbalanced. Therefore, we train the classification model after applying SMOTE algorithm to datasets. SMOTE is an oversampling technique by synthetizing new samples from the minority [3].

Algorithm 3. SGD for training the FM model

input: training set $\boldsymbol{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_m, y_m)\}$, learning rates η
output: the parameters of FM model, $\Theta = (w_0, \boldsymbol{w}, \boldsymbol{v})$
initialization: $w_0 := 0; \boldsymbol{w} := 0; \boldsymbol{v} \sim \mathcal{N}(0, \sigma)$
for $(x_i, y_i) \in \boldsymbol{D}$ **do**
  $w_0 = w_0 - \eta(\frac{\partial loss(\hat{y}(\boldsymbol{x}_i), y_i)}{\partial w_0} + 2\lambda^0 w_0)$
  for $i \in \{1, 2, ..., n\}$ **do**
    if $\boldsymbol{x_i} \neq 0$
      $w_i = w_i - \eta(\frac{\partial loss(\hat{y}(\boldsymbol{x}_i), y_i)}{\partial w_i} + 2\lambda^{w_{\pi(i)} w_i})$
      for $j \in \{1, 2, ..., k\}$ **do**
        $v_{ij} = v_{ij} - \eta(\frac{\partial loss(\hat{y}(\boldsymbol{x}_i), y_i)}{\partial v_{ij}} + 2\lambda^v_{\pi(i),j} v_{ij})$
stop until meeting criterion

There are 660 users distributing in 4 SELs in this research. We implement the FM-based classification through a 5-fold cross-validation over the training set. The normalized confusion matrix is shown as Fig. 3. The model achieves the accuracy over 85% when predicting the users in SEL B. However, it performs badly in identifying the users from other classes. Most users from SEL A and SEL C are mistakenly predicted as SEL B. This phenomenon is caused by the fact that the users from SEL B occupy over 75%. Thus, the model learns the features of users in SEL B better. For the classes with less users, the model is not able to learn enough information to predict them accurately. Therefore, the predictive model is inclined to regard users without obvious characteristics as SEL B.

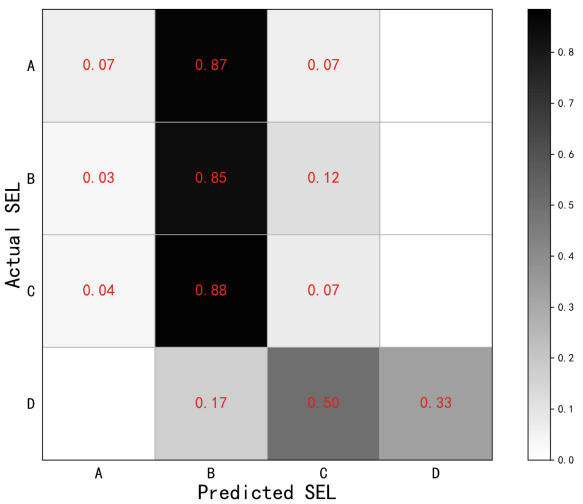

Fig. 3. Normalized confusion matrix.

In order to solve the class-imbalance problem, we apply the SMOTE algorithm to the row datasets. The number of users distributed in 4 classes is same after the SMOTE oversampling. The normalized confusion matrix based on SMOTE is shown as Fig. 4. The accuracies of predicting the users in SEL A, SEL C and SEL D are extremely improved.

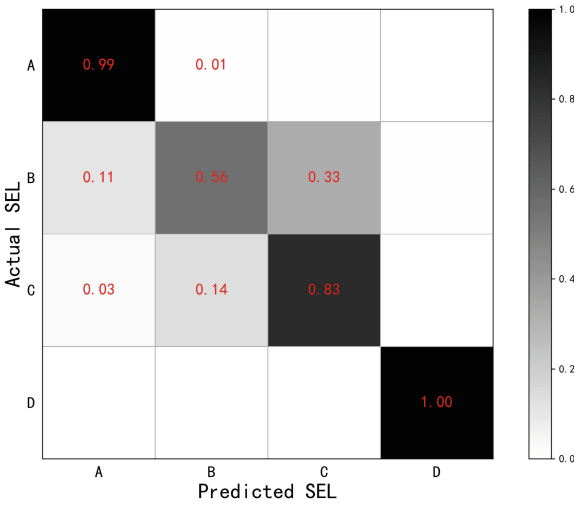

Fig. 4. Normalized confusion matrix based on SMOTE.

The FM-based model performs well on the datasets after oversampling. We implement the predictive model through a 5-fold cross-validation over the training set and achieve the accuracy of 84.3%. In addition, we train predictive models with Random Forests (RF) and SVM for comparison. Table 5 shows the precision, recall, F1 score and accuracy of different models. Our model performs better than RF and SVM.

Table 5. The examples of app usage records.

Model	Precision	Recall	F1	Accuracy
Random forests	0.59	0.68	0.63	0.680
SVM	0.58	0.76	0.65	0.756
FM-based model	0.84	0.84	0.84	0.843

SGD is used to learn the parameters of FM model. In order to investigate how the SGD influences the accuracy of our model, we compute the accuracy versus regularization coefficient λ as shown in Fig. 5. With the decrease of λ,

the accuracy is significantly improved. The larger λ is, the more serious punishment to the model. Thus, the model is too simple to learn the features of samples. However, it is important to learn the information from each sample for classification model. Therefore, we set a relatively small λ.

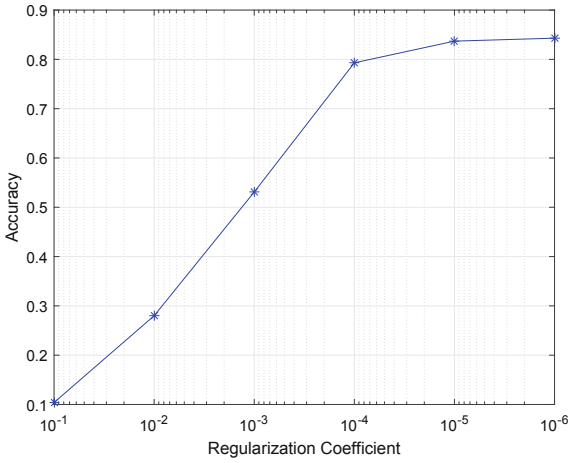

Fig. 5. Accuracy versus regularization coefficient λ.

In addition, we compute the accuracy versus learning rate α as shown in Fig. 6. With the increase of α, the accuracy increases gradually. While the accuracy becomes decreasing when α is over 0.8.

6 Related Work

Inferring the SELs of individuals is important for both policy makers and companies. The traditional approaches to predicting the individuals' economic status mainly rely on census. With the development of data mining technologies, many researchers attempt to predict SELs of individuals via novel datasets and algorithms.

Some researchers found that there exist correlations between social networks and the economic status of individuals [1,2,16]. Shaojun Luo et al. extracted social networks from user's calling data to infer individuals' SELs [14]. The majority of researchers focus on the research of using mobile phone data to estimate individuals' economic status. Luca Pappalardo et al. studied the links between human mobility extracted from calling data and economic development [18]. Other studies employed calling data to investigate consumer behaviors [4] and economic development [5,8].

Other researchers attempt to use data collected from Internet to discover indicators about society or economics. Michal Kosinski et al. [15] predicted users'

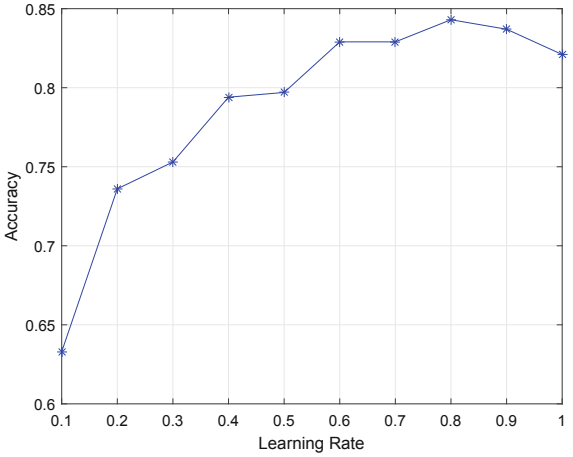

Fig. 6. Accuracy versus learning rate α.

personal attributes like intelligence, ethnicity from digital records of Facebook. Renato Miranda Filho et al. [6] took advantage of user data obtained from Foursquare and Twitter to propose a method of inferring user social class. Other researchers analysed user occupation class through the social content of Twitter [20].

7 Conclusions

This research proposes a method to predict SELs of individuals using their App usage records provided by China Telecom. We first define the SEL according to the social stratification method in sociology. With a dataset containing 660 users provided by the operator China Telecom, we divided 660 users into 4 classes. We then train the FM-based model from the datasets after SMOTE oversampling. The accuracy achieves 84.3%, which performs much better than the previous research using tweets [6].

Although this is a pioneering work using App usage records to predict socio-economic status, there are still some limitations. The samples in our study only cover 4 socio-economic classes except the lower socio-economic level. In addition, our method may not work in the environment where smart phones and Apps are not popular. However, this paper proved the correlations between users' App usage behaviors and their socio-economic status, which provides a promising method and idea for other researchers.

Acknowledgements. The work is supported in part by Science, Technology and Innovation Commission of Shenzhen Municipality (No. JCYJ20170816151823313), NSFC (No. U1711263, 61501527), States Key Project of Research and Development Plan (No. 2017YFE0121300-6), Fundamental Research Funds for the Central Universities,

MOE-CMCC Joint Research Fund of China (MCM20160101), Guangdong Science and Technology Project (No. 2016B010126003) and Guangdong Provincial Special Fund For Modern Agriculture Industry Technology Innovation Teams (No. 2019KJ122).

References

1. Arveson, W.: Methods and Applications (2002)
2. Caldarelli, G., Vespignani, A.: Large Scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Science. World Scientific Publishing Co., Inc., River Edge (2007)
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**(1), 321–357 (2002)
4. Decuyper, A., et al.: Estimating food consumption and poverty indices with mobile phone data. CoRR, abs/1412.2595 (2014). http://arxiv.org/abs/1412.2595
5. Eagle, N., Macy, M., Claxton, R.: Network diversity and economic development. Science **328**(5981), 1029–1031 (2010). https://doi.org/10.1126/science.1186605
6. Filho, R.M., Borges, G.R., Almeida, J.M., Pappa, G.L.: Inferring user social class in online social networks (2014)
7. Freeman, L.C.: Centrality in social networks conceptual clarification. Soc. Netw. **1**(3), 215–239 (1978)
8. Gutierrez, T., Krings, G., Blondel, V.D.: Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. CoRR, abs/1309.4496 (2013). http://arxiv.org/abs/1309.4496
9. Huang, Y., Yu, L., Xiang, W., Cui, B.: A multi-source integration framework for user occupation inference in social media systems. World Wide Web-internet Web Inf. Syst. **18**(5), 1247–1267 (2015)
10. Lampos, V., Aletras, N., Geyti, J.K., Zou, B., Cox, I.J.: Inferring the socioeconomic status of social media users based on behaviour and language. In: European Conference on Information Retrieval (2016)
11. Li, C.: Prestige stratification in contemporary chinese society. Sociol. Stud. **2**, 74–102 (2005)
12. Li, H., et al.: Characterizing smartphone usage patterns from millions of android users. In: Internet Measurement Conference (2015)
13. Lindén, A.L.: Travel patterns and environmental effects now and in the future: implications of differences in energy consumption among socio-economic groups. Ecol. Econ. **30**(3), 405–417 (1999)
14. Luo, S., Morone, F., Sarraute, C., Travizano, M., Makse, H.A.: Inferring personal economic status from social network location. Nature Commun. **8**, 15227 (2017)
15. Michal, K., David, S., Thore, G.: Private traits and attributes are predictable from digital records of human behavior. PNAS **110**(15), 5802–5805 (2013)
16. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45**(2), 167–256 (2003)
17. Pan, W., Ghoshal, G., Krumme, C., Cebrian, M., Pentland, A.: Urban characteristics attributable to density-driven tie formation. Nature Commun. **4**(3), 1961 (2012)
18. 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)

19. Pappalardo, L., Vanhoof, M., Gabrielli, L., Smoreda, Z., Pedreschi, D., Giannotti, F.: An analytical framework to nowcast well-being using mobile phone data. Int. J. Data Sci. Anal. **2**(1–2), 1–18 (2016)
20. Preotiucpietro, D., Lampos, V., Aletras, N.: An analysis of the user occupational class through twitter content (2015)
21. Preot, D., Lampos, V., Aletras, N.: An analysis of the user occupational class through twitter content (2015)
22. Propper, C., Damiani, M., Leckie, G., Dixon, J.: Impact of patients' socioeconomic status on the distance travelled for hospital admission in the english national health service. J. Health Serv. Res. Policy **12**(3), 153–159 (2007)
23. Qiang Li, H.L.: Vocational prestige in transition. Acad. Res. **12**, 34–42 (2009)
24. Rendle, S.: Factorization machines. In: IEEE International Conference on Data Mining (2011)
25. Van Canneyt, S., Bron, M., Haines, A., Lalmas, M.: Describing patterns and disruptions in large scale mobile app usage data. In: Proceedings of the 26th International Conference on World Wide Web Companion, WWW 2017 Companion, pp. 1579–1584 (2017). https://doi.org/10.1145/3041021.3051113
26. Welke, P., Andone, I., Blaszkiewicz, K., Markowetz, A.: Differentiating smartphone users by app usage. In: ACM International Joint Conference on Pervasive & Ubiquitous Computing (2016)
27. Ye, X., et al.: Preference, context and communities:a multi-faceted approach to predicting smartphone app usage patterns. In: International Symposium on Wearable Computers (2013)