Utilizing Geospatial Information in Cellular Data Usage for Key Location Prediction

Yinan Yu
School of Business, The University of Hong Kong
yyn10695@connect.hku.hk

Baojun Ma
School of Economics and Management, Beijing University of Posts and Telecommunications
mabaojun@bupt.edu.cn

Hailiang Chen
Department of Information Systems, City Univerisity of Hong Kong
hailchen@cityu.edu.hk

Benjamin Yen
School of Business, The University of Hong Kong
benyen@business.hku.hk

Abstract

Previous research on the identification of key locations (e.g., home and workplace) for a user largely relies on call detail records (CDRs). Recently, cellular data usage (i.e., mobile internet) is growing rapidly and offers fine-grained insights into various human behavior patterns. In this study, we introduce a novel dataset containing both voice and mobile data usage records of mobile users. We then construct a new feature based on the geospatial distribution of cell towers connected by mobile users and employ bivariate kernel density estimation to help predict users' key locations. The evaluation results suggest that augmented features based on both voice and mobile data usage improve the prediction precision and recall.

1. Introduction

Mobile technology continues to scale rapidly and cellular data consumption is showing promising growth over the years. According to a report from PwC Communications Review (2014), people spend 84% of the time on mobile Internet when they use their phones, which dwarfs the time of making phone calls [1]. The white paper of Cisco Visual Networking Index also indicates that global mobile data traffic grew 74 percent, and the mobile data revenue of carriers eclipsed voice revenue in 2015 [2]. The advancement of mobile internet technology provides a novel source of massive data recording whereabouts of people in space and time. Although wireless carriers' transaction logs are a timely and cost-efficient data source, they may not contain the home or workplace address for every user, which could be an important information item for customer profiling. In the business analytics era, for business applications such as precision marketing, one of the first steps is to know where a customer lives or works. Based on such information, businesses can offer customized advertisements and location-based services to their customers. For research in transportation and urban planning, scholars also need to identify meaningful locations that serve as reference points for analyzing people's travel behavior and mobility patterns.

Previous studies have mainly relied on call detail records (CDRs) to develop algorithms that help identify meaningful places (e.g., home and workplace) for different individuals [3-5]. CDRs provide a cost-effective alternative to overcome the drawbacks of traditional travel surveys, such as small sample sizes and long time intervals. They have also been used to investigate other issues like socioeconomic characteristics prediction [6] and disease transmission prediction [7, 8]. However, cellular data consumption is rarely used in the existing literature in spite of its astonishing growth, which may be due to the lack of data available to researchers. One exception in IS area is the study by Ghose and Han (2011) [9]. They investigate what factors affect individual's mobile internet content generation and consumption. Moreover, prior algorithms on key location prediction only utilize the counting information of events that arise from different cell phone usage behavior. Information on the spatial distribution of locations visited by mobile users is underexploited. In this study, we describe an approach to predict key locations of human activities by taking advantage of a novel dataset containing both voice call and cellular data usage records. In particular, we leverage users' two-dimensional location information recorded by the log of transactions between users and cell towers. By incorporating both the spatial location and temporal frequency of mobile usage into our proposed algorithm, we expect it could potentially improve the prediction accuracy of identifying key locations.

We obtain a one-month dataset of mobile phone transaction records for a random sample of users in one of the largest cities in China from one mobile operator. The dataset contains both CDRs and mobile data usage

URI: http://hdl.handle.net/10125/50010
ISBN: 978-0-9981331-1-9

information. The company also provides us with anonymized billing addresses of either home or workplace for our sampled users. To protect the confidentiality of sampled users, the locations of the cell towers surrounding a user's billing address within a small radius are used as proxies for the user's home or workplace location. We propose a new algorithm for home/workplace prediction based on the spatial distribution of cell towers connected by users over a time period. For each user, we place a kernel (probability density) on the location of each used tower, weight each tower by a regularity measure (e.g., its intensity of being connected by the user in a regular pattern), and then use kernel density estimation to obtain the distribution of density in its surrounding area. The density at any location is an estimate of its probability of being the focal user's home or workplace location. After aggregating the densities of all kernels, we get a smoothed bivariate probability density in the whole study area. We assume that the point with the maximum estimated kernel density is the place where the address of interest is most likely located in. To incorporate this information into the prediction model, we propose a new feature to measure the distance between each tower and the point with maximum kernel density.

Our study does not aim to infer the exact location where people live or work, but to discover the approximate meaningful places where people spend a huge amount of their time. These key locations serve as the anchor points to study people's daily activities, mobility, and other behavior patterns [5][10]. We evaluate the predictive power of our augmented features (e.g., geospatial distance and mobile usage behavior) on the extended dataset (e.g., CDR and cellular data usage records), and compare their performance with prediction models based on voice records only. Our results demonstrate that (1) information from cellular data usage can help improve the prediction accuracy in identifying the home and workplace locations; and (2) the geospatial information revealed from cellular data usage provides additional value in understanding mobile users' behavioral patterns.

2. Related Literature

The design of algorithms on users' home/workplace prediction in previous works is mainly based on individuals' cell phone calling behavior. The fundamental idea is that people spend a large amount of time in meaningful locations like home and workplace regularly. Based on this concept of regularity, some studies compute the number of days a user connects to different towers [3][5], and consider the tower with the highest regularity value as the location of home. Other studies take a step further to analyze people's different calling behaviors at different locations. For example, Ahas et al. (2010) compute the average and standard deviation of the start times of all phone calls for each user [3]. They explain that the average start times of workplace activities should begin at working hours. Additionally, people undertake a higher variety of activities at home than at workplace, so the standard deviation of connections' starts times should show different patterns. Algorithms are also developed from the perspective of inactivity [4]. Inactivity is defined as an event with the time difference between two consecutive transactions exceeding a threshold, which aims to model human's resting behavior. Generally, the tower located in the home area is the one with highest inactivity frequency.

Most of these studies rely on counting different events that potentially reflect people's mobile usage behavior. However, the location information of connected cell towers is largely ignored. From a geospatial perspective, the problem of key location identification is equivalent to predicting the probability of each tower being in the home/workplace area. The spatial distribution of towers connected by a user contains important information to assess such probabilities. One main purpose of this study is to utilize this information to develop a new prediction feature.

3. Methodology

3.1. Mobile Phone Dataset

The anonymous mobile phone dataset is obtained from one of the largest telecommunication companies in China. It contains log records of both voice and 3G data usage for a random sample of users in one of the largest cities in the country in April 2014. These records are communication transactions between mobile devices and base transceiver stations (BTS) of the mobile operator. Each time a user calls or consumes 3G data, the mobile operator registers the nearest available cell tower to the user, and the system records user ID, starting time and duration of this transaction, traffic of data consumed, GSM cell tower ID, and location area code (LAC) of the tower. Such information allows us to locate the user at the resolution of the connected tower's coverage area. The dataset also provides users' basic demographic information such as age and sex.

We obtain the BTS dataset updated to the end of 2015. This dataset provides information on all the cell towers of the mobile operator in the focal city, consisting of cell ID, LAC, location, latitude, and longitude of each tower. We merge the BTS dataset with the mobile phone dataset to obtain the locations where cellular users get connected with towers.

In order to protect the privacy and confidentiality of its users, the telecom company does not disclose users' billing addresses to us directly. It conducts a series of processes to clean and remove personally identifiable information from the data. First, users reporting meaningless billing addresses, addresses at a quite coarse granularity, or no address at all are screened out. Second, users are split into two groups with either a home or a workplace address, according to the location of the billing address reported. Then all townships of the focal city are grouped into three clusters by the distribution of tower density (the ratio of the number of cell towers to the area of a township). Different cutoff values of tower coverage are set for these three groups given that more populated areas tend to have more towers. The townships with the highest tower density (e.g., tower density ⩾ 1000 per km^2) are urban central areas and are assigned the cutoff value of 150 meters. The cutoff values of the second group (e.g., 200 per km^2 ⩽ tower density < 1000 per km^2) and the third group (e.g., tower density < 200 per km^2) are set to be 300 meters and 2000 meters, respectively. Towers located within the coverage radius (the cutoff value) of the address in the corresponding township are considered potential target towers that people connect to when they are at home or workplace.

Our sample contains 4,176 users who generate over 3.8 million cellular data transactions and 0.8 million voice transactions in a month. These users consume both voice and data during our study period, so that prediction performances of different datasets generated by the same set of individuals can be compared. There are 913 data transactions (30 per day) and 208 call transactions (7 per day) for each user on average. Each user consumes 373.15 MB cellular data on average in that month.

Our dataset has some advantages compared with those used in previous studies. Prior studies on algorithms for identification of meaningful locations usually use call detail records. However, as we mentioned above, cellular data consumption has witnessed rampant growth that overshadows voice call and text messaging [1]. With the established fact that more people tend to use data more frequently compared with voice, it may cause some prediction bias if we only use CDRs. Our dataset contains both voice call and 3G data usage information, which allows us to assess the prediction performance of each type of data and develop new algorithms and features to capture the unique characteristics of cellular data. Second, most of the prior studies do not have the home/workplace locations reported by users. Some studies validate the accuracy of their algorithms by comparing the population distribution (or job sector distribution) with the distribution of predicted home (or workplace) locations aggregated at the planning area level [3, 4]. The availability of user reported addresses (although anonymized for privacy reasons in our context) can reduce bias and ensure accuracy of our prediction model.

Following prior algorithms based on CDRs we construct the following features. For each individual in our sample and for each cell tower he/she has connected to at least once, we count the number of days in which user *i* sends at least one request to tower *j* as the measure of regularity (i.e., $Regularity_{ij}$). We also construct regularity measures based on weekday and weekend (including public holidays as well) counts for workplace prediction (i.e., $RegWeekday_{ij}$ and $RegWeekend_{ij}$) respectively, because people usually do not go to their workplace on weekends and public holidays. Users who connect to their most frequently used tower in less than 7 days are screened out, since such kind of users do not exhibit enough regularity in their voice or data usage behavior. We calculate the average start time of all connections for each tower, and create a dummy variable to indicate whether the average start time begins at working or non-working hour (i.e., $ConnectAtWorkhour_{ij}$). The standard deviation of start times for the transactions of each tower (i.e., $StdConnHour_{ij}$) is also computed. For each pair of consecutive transactions, we compute the time difference (or lag) between them to gauge the inactivity of a tower connected by a user. If the lag exceeds a threshold (e.g., 5 hours), the number of inactivities increments by one. $Inactivity_{ij}$ is the count of inactivities divided by regularity to cope with the imbalance in inactivity among users with different regularity. Each of these features is calculated in two different datasets (i.e., dataset containing CDRs only and dataset combining voice and cellular data usage together). All the continuous variables are Z-score normalized.

3.2. Kernel Density Estimation

Kernel density estimation (KDE) is a data smoothing method to calculate the probability density of the neighborhood area of observation points [11, 12]. Each observation point in the sample is located on a two-dimensional surface and is allocated a kernel, namely probability density. Each observation is overlaid by an area, the size of which is determined by bandwidth parameter. And then the probability density of the area is estimated by using a certain kernel function. Probability density of the intersection of different areas is the overlap of densities of all kernels superposing that point. Densities of different observation points can also be weighted by a certain measure of interest. For the point of evaluation in the study area, observations located close to it and with higher weights contribute more to its estimation. Two-dimensional kernel density estimator at *x* is defined as in Equation (1). X_i denotes vectors of x-y coordinates which describe the location of observations, and *n* is the number of observations. *x* is a vector of x-y coordinates which describe the location of the grid where the function is being estimated. K (•) is the kernel function, which defines how each observation contributes differently to the density estimation of area x based on its proximity. *h* is the bandwidth restricting the search radius. The narrower the bandwidth is, the greater influences nearby observations contribute.

$$\hat{K}(x) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left\{\frac{x-X_i}{h}\right\} \qquad (1)$$

The observation points in our context are cell towers connected by users during the study period. All towers are located in the x-y projection coordinates of their latitudes and longitudes. Each cell tower is weighted by different regularity measures depending on whether the task is to identify a home or a workplace location. For the group of users reporting residential addresses, towers are weighted by $Regularity_{ij}$; for the group of users reporting workplace addresses, towers are weighted by the ratio $(RegWeekday_{ij}+1)/(RegWeekend_{ij}+1)$ (constant 1 is added to the denominator to avoid division by zero condition). We use the Gaussian kernel function, one of the most widely used functions in KDE, to estimate the probability density of towers appearing in its local neighborhood with a search radius of 1 kilometer. The choice of this bandwidth parameter stems from actual tower coverage. Then all local densities are aggregated to yield an overall density for each user. Figure 1 and Figure 2 are three-dimensional and two-dimensional schematic diagrams of bivariate kernel density estimated by using cellular data usage of one user in our sample. In Figure 1, the red dots are cell towers connected by this user, located on a surface with x-y coordinates. X-coordinate and Y-coordinate are projection coordinates of the latitude and longitude of the observation points. The height of the three-dimensional shape is the value of kernel density estimated after overlapping the density of each kernel. Figure 2 is the overhead view. Again, dots are towers connected by this user. Different colors represent different values of kernel density, with warm colors (e.g., red and yellow) indicating higher values and cold colors (e.g., green and blue) indicating lower values. The size of the tiny grid is our display resolution (100 meters × 100 meters). Solid grey lines in the figure denote the boundaries of townships. We assume that the point (geometric center of the grid) with the maximum kernel density is the location where the target address (home/workplace) is most likely located in. The new feature we propose, $KDEDistance_{ij}$, is the distance between each tower and this point, which is negatively correlated with the probability of being the target user address.

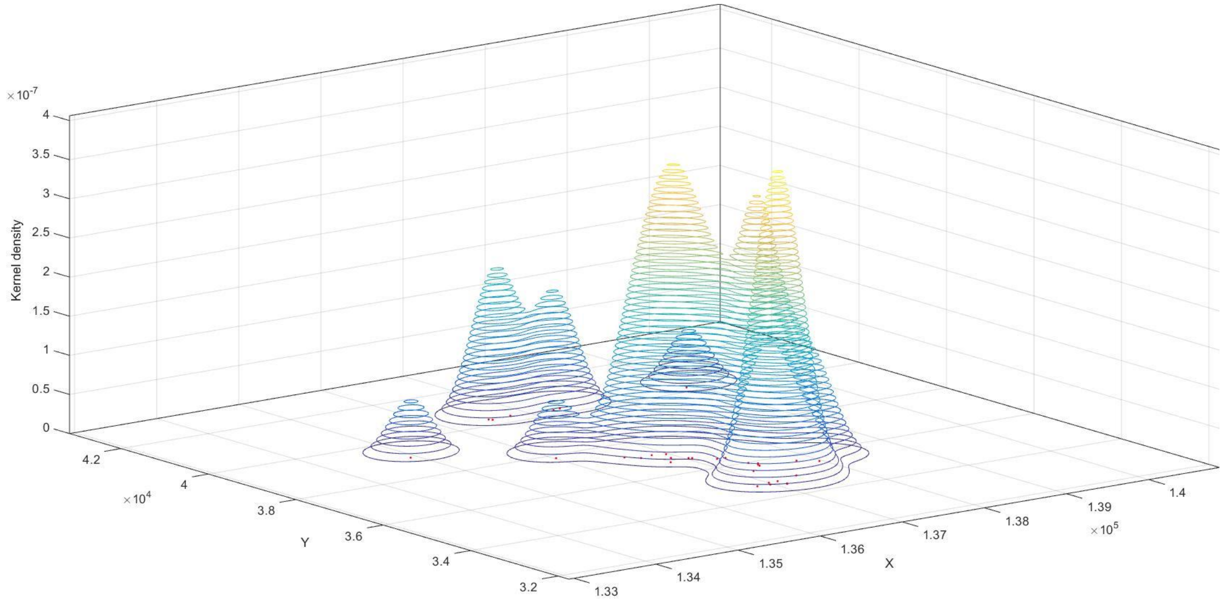

Figure 1. Three-dimensional schematic diagram of kernel density

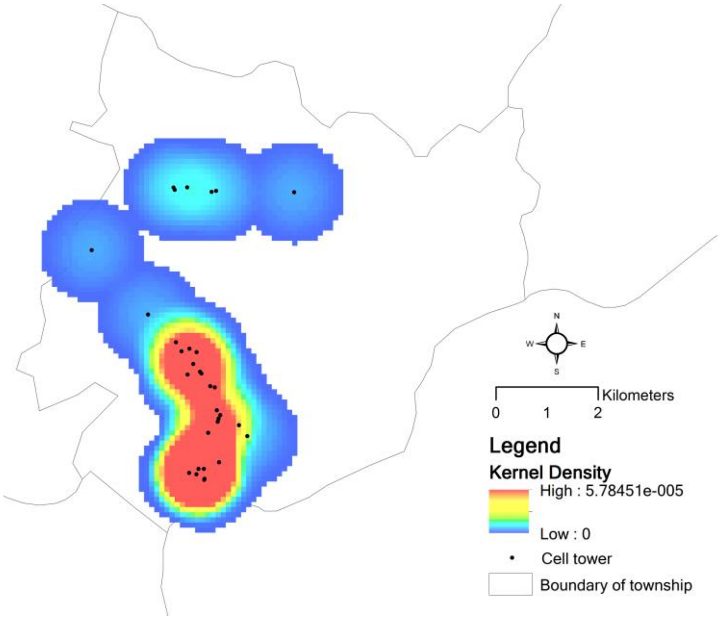

Figure 2. Two-dimensional schematic diagram of kernel density

3.3. Prediction Model

$$\Pr\left(y_{ij}=1\middle|X_{ij}\right)=\frac{\exp\left(X_{ij}^{'}\beta\right)}{\exp\left(X_{ij}^{'}\beta\right)+1} \qquad (2)$$

We use Probit model to predict the probability of tower *j* being a home/workplace tower for user *i* (Equation (2)). The unit of analysis is at the user-tower level, meaning that each record is a tower ever connected by a certain user within our study period. For each user *i*, we have a binary outcome y_{ij} for each tower *j*. For the group of users with a residential address we calculate the probability of whether the tower he/she used is a home tower or not (H_{ij}). For the group of users with a workplace address, y_{ij} denotes the probability of being a workplace tower (W_{ij}). X'_{ij} denotes the features described above: *KDEDistance*$_{ij}$, *Regularity*$_{ij}$ (or *RegWeekday*$_{ij}$ for the dependent variable W_{ij}), *ConnectAtWorkhour*$_{ij}$, *StdConnHour*$_{ij}$, and *Inactivity*$_{ij}$. We run the regression on two datasets: one contains voice records only and the other consists of both voice and data usage records.

4. Evaluation

In this section, we evaluate (1) the contribution of the dataset combining voice and data usage records beyond that of voice records only; and (2) whether our proposed feature based on the geospatial distribution of connected towers, *KDEDistance*$_{ij}$, improves key location prediction. The entire sample is split into a training dataset (70% of the users) and a holdout dataset (30% of the users). We first use the training dataset to train a binary classifier, and then test it on the holdout dataset. In each dataset, users are divided into two groups according to their address types: home and workplace.

Regarding the contribution of data usage records, results in Table 1 show that prediction models perform better using the dataset containing both voice and data usage records than using voice records only. We measure two evaluation metrics, precision and recall, for both positive and negative classes. We define target towers, those located within the cutoff values of users' addresses, as the positive (+) class and non-target towers as the negative (-) class. The threshold values of predicted probability for the positive class are chosen so that the number of predicted home/workplace towers is roughly consistent with the real home/workplace tower distribution (e.g., the ratio of home/workplace towers to all towers). For home prediction, models using the combined dataset outperform the models based on the voice dataset in every performance measure. For workplace prediction, only the precision for the positive class and the recall for the negative class on holdout data do not show improved performance. Since the unit of analysis is at the user-tower level in Table 1, we further report another set of performance measures at the user level. Table 2 shows that usage of the combined dataset improves the user level prediction accuracy too.

Table 1. User-tower level results based on different datasets

	Precision				Recall			
	Training		Holdout		Training		Holdout	
	+	-	+	-	+	-	+	-
Home Prediction								
CDR	0.4174	0.8460	0.4640	0.8376	0.4178	0.8458	0.4629	0.8382
CDR + Data	**0.4365**	**0.8649**	**0.5255**	**0.8675**	**0.4469**	**0.8599**	**0.4652**	**0.8676**
Workplace Prediction								
CDR	0.4325	0.8169	0.4203	0.8414	0.4324	0.8170	0.4191	0.8420
CDR + Data	**0.4375**	**0.8272**	0.3903	**0.8742**	**0.4373**	**0.8273**	**0.5810**	0.7624

Table 2. User level results based on different datasets

	Precision				Recall			
	Training		Holdout		Training		Holdout	
	+	-	+	-	+	-	+	-
Home Prediction								
CDR	0.6395	0.9631	0.6211	0.9563	0.6395	0.9918	0.6211	0.9871
CDR + Data	**0.7308**	**0.9790**	**0.8000**	**0.9760**	**0.7308**	**0.9964**	**0.8000**	**1.0000**
Workplace Prediction								
CDR	0.6028	0.9464	0.5992	0.9817	0.6028	0.9892	0.5967	0.9871
CDR + Data	**0.7235**	**0.9792**	**0.7949**	**0.9847**	**0.7235**	**0.9930**	**0.7908**	**1.0000**

Next, in Table 3 and Table 4, we demonstrate the effectiveness of the new feature based on the geospatial distribution information of connected towers by a user. Table 3 presents the regression results of the Probit model. Column (1) and (2) show the results for home prediction, and Column (3) and (4) are for workplace prediction. Column (2) and (4) are models including the newly proposed feature, $KDEDistance_{ij}$. The coefficients on $KDEDistance_{ij}$ are significantly negative, which indicates that the farther the tower is located from the point with maximum kernel density, the lower its probability of being in the home/workplace area. Additionally, after including $KDEDistance_{ij}$ in the model, Pseudo R^2 is largely improved in both prediction conditions.

Table 4 presents the prediction results on two feature sets: features based on counting information only (i.e., without $KDEDistance_{ij}$) and all features. Generally, our prediction model performs better with the new feature for both home and workplace predictions. Specifically, the precision and recall for the positive class improve a lot, and the precision and recall for the negative class also show marginal improvements. The only exception is the recall for the negative class in workplace prediction on the holdout dataset, for which the new feature does not show prediction improvement.

Table 3. Probit regressions on two feature sets

VARIABLES	Home		Workplace	
	(1)	(2)	(3)	(4)
$KDEDistance_{ij}$		-0.785***		-0.638***
		(0.025)		(0.013)
$Regularity_{ij}$	0.099***	0.069***		
	(0.012)	(0.013)		
$RegWeekday_{ij}$			0.095***	-0.046**
			(0.021)	(0.022)
$ConnectAtWorkhour_{ij}$	-0.189***	-0.133***	-0.158***	-0.110***
	(0.019)	(0.020)	(0.015)	(0.016)
$StdConnHour_{ij}$	0.050***	0.038***	0.121***	0.078***
	(0.01)	(0.010)	(0.007)	(0.007)
$Inactivity_{ij}$	0.160***	0.073***	0.097***	0.160***
	(0.014)	(0.015)	(0.021)	(0.022)
Constant	-0.938***	-1.132***	-0.647***	-0.809***
	(0.010)	(0.015)	(0.011)	(0.012)
Pseudo R^2	0.037	0.142	0.028	0.110
#Users	286	286	434	434
#Observations	30,295	30,295	38,108	38,108

Table 4. Results based on different features

	Precision				Recall			
	Training		Holdout		Training		Holdout	
	+	-	+	-	+	-	+	-
Home Prediction								
Without *KDEDistance*	0.3273	0.8375	0.3908	0.8296	0.3349	0.8328	0.3905	0.8298
All features	**0.4365**	**0.8649**	**0.5255**	**0.8675**	**0.4469**	**0.8599**	**0.4652**	**0.8676**
Workplace Prediction								
Without *KDEDistance*	0.3649	0.8048	0.3279	0.8240	0.3642	0.8053	0.3276	0.8242
All features	**0.4375**	**0.8272**	**0.3903**	**0.8742**	**0.4373**	**0.8273**	**0.5810**	0.7624

5. Conclusion and Future Work

In this study, we predict key locations of human activities using a comprehensive mobile phone dataset and propose a feature based on the geospatial information of towers connected by mobile users. Our dataset consists of both phone call records and 3G data usage records, filling the gap that cellular data consumption information has largely not been used in academic research. Prior studies also often neglect the spatial distribution of users' whereabouts recorded in mobile transaction logs. The new feature we construct is the distance between each tower and the point with the maximum kernel density of being a user's home or workplace location. Our evaluations show that adding cellular data usage information is effective in improving the precision and recall rates of identifying the home/workplace for mobile users. In addition, our results confirm the importance of considering geospatial information when predicting key locations.

As an ongoing research, this study has some limitations. First, the choice of the bandwidth parameter in kernel density estimation is based on tower coverage. Although determining the optimal bandwidth of a bivariate KDE is still an open question [13], we have tested other choices for more robustness checks. Second, we currently use the Probit regression model as our binary classifier. Other classifiers such as Support Vector Machine and Artificial Neural Networks can also be adopted. Third, our data are split into a training dataset and a holdout dataset. K-fold cross validation can be conducted to reduce overfitting and increase generalizability. Finally, other algorithms can be developed to further utilize the information revealed from cellular data consumption.

6. Acknowledgements

The author Baojun Ma (corresponding author) is supported by the National Natural Science Foundation of China (No. 71772017 / 71402007) and the Beijing Municipal Social Science Foundation (No. 17GLB009).

7. References

[1] PwC Communication Reviews. 2014. "Mobile Data Analytics: Not Just for Consumers Any More," http://www.pwc.com/gx/en/industries/communications/publications/communications-review/mobile-data-analytics.html

[2] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020 White Paper. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[3] Ahas, R., Silm, S., Jarv, O., Saluveer, E., and Tiru, M. 2010. "Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones," Journal of Urban Technology (17:1), pp. 3-27.

[4] Dash, M., Nguyen, H.L., Hong, C., Yap G.E., Nguyen, M.N., Li, X., Krishnaswamy S.P., Decraene, J., Antonatos, S., Wang Y., Anh, D.T., Shi-Nash, A. 2014. "Home and Work Place Prediction for Urban Planning Using Mobile Network Data," IEEE 15th International Conference on on Mobile Data Management, Volume: 2.

[5] Xu, Y., Shaw, S., Zhao, Z., Yin, L., Fang, Z., and Li, Q. 2015. "Understanding Aggregate Human Mobility Patterns Using Passive Mobile Phone Location Data: a Home-based Approach. Transportation (42), pp. 625–646.

[6] Blumenstock, J., Cadamuro, G., and On, R. 2015. "Predicting Poverty and Wealth From Mobile Phone Metadata. Science (350:6264), pp. 1073-1076.

[7] Bengtsson, L., Gaudart, J., Lu, X., Moore S., Wetter, E., Sallah, K., Rebaudet, S., and Piarroux, R. 2015. "Using Mobile Phone Data to Predict the Spatial Spread of Cholera," Scientific Reports, 5 (8923).

[8] Wesolowski, A., Metcalf, C.J.E., Eagle, N., Kombich, J., Grenfell, B.T., Bjørnstad, O.N., Lessler, J., Andrew J. Tatem, and Buckee, C.O. 2015. "Quantifying Seasonal Population Fluxes Driving Rubella Transmission Dynamics Using Mobile Phone Data," Proceedings of the National Academy of Science of the United States of America (112:35), pp. 11114–11119.

[9] Ghose, A., Han, S.P. 2011. "An Empirical Analysis of User Content Generation and Usage Behavior on the Mobile Internet," Management Science (57:9), pp. 1671–1691.

[10] Gonzalez, M. C., Hidalgo, C.A., and Barabasi, A. 2008. "Understanding Individual Human Mobility Patterns," Nature (453), pp. 779–782.

[11] Silverman, B. W. Density Estimation for Statistics and Data Analysis. New York: Chapman and Hall, 1986.

[12] Seaman, E., Powell, R.A. 1996. "An Evaluation of the Accuracy of Kernel Density Estimators for Home Range Analysis," Ecology (77:7), pp. 2075-2085.

[13] Kohler, M., Schindler A., and Sperlich, S. 2014. "A Review and Comparison of Bandwidth Selection Methods for Kernel Regression," International Statistical Review (82:2), pp. 243–274.