Home and Work Place Prediction for Urban Planning Using Mobile Network Data

Manoranjan Dash, Hai Long Nguyen, Cao Hong,
Ghim Eng Yap, Minh Nhut Nguyen, Xiaoli Li,
Shonali Priyadarsini Krishnaswamy
Institute of Infocomm Research
*A*Star, Singapore 138632*
{*dashm, nguyenhl, hcao, geyap, mnnguyen, xlli, spkrishna*}
@i2r.a-star.edu.sg

James Decraene, Spiros Antonatos,
Yue Wang, Dang The Anh,
Amy Shi-Nash
R&D Labs, Living Analytics, Group Digital Life,
Singapore Telecommunication Limited, Singapore
{*jdecraene, antonatos, wangyue, anhkeen,*
amyshinash}*@singtel.com*

***Abstract*—We present methods to predict and validate home and work places of anonymized users using their mobile network data. Knowledge of home and work place of a user is essential in order to find his (and overall population) mobility profiles. There are many methods that predict home and work places using GPS data. But unlike GPS data, mobile network data using GSM do not provide the exact location of a phone event. We use a novel criterion that combines an extracted feature from mobile data (i.e., inactivity – no phone event for a given period of time) with open source data about location category to predict home location. Results show that the new criterion gives better prediction accuracy than inactivity alone. We predict work place using the idea that one goes to her work place on most of the weekdays but rarely on weekends. We validate our methods by comparing against the ground truth obtained from open source data. Validation results show that our proposed methods are about 25% more accurate than existing methods both for home and work place predictions.**

***Keywords*-Mobile network data; Home and Work Place Prediction; Urban Planning;**

I. Introduction

Mobile network data, if analyzed properly, can increase our knowledge about mobility profiles of people in a place. Such knowledge can be used in many applications such as product advertisement, traffic management. A critical constant feature of any mobility profile is the knowledge of home and work places. Home and work place distribution of a city also helps in making urban development decisions. However, such data would typically be collected via surveys and thus be limited in size. In this work, we describe methods that are used for predicting and validating home and work places using large mobile network data.

For this research project, a sample mobile network data of three months is used. Mobile network data is the service log when a mobile phone is connected to mobile network. It contains anonymised ID, latitude, longitude, time stamp and service type (i.e. voice, SMS and data records). The anonymised ID is a machine generated ID via a two-step non-reversible AES encryption and hash process. This means it is not possible to trace back to the original ID. There is no personal information about mobile subscribers in the data set, nor any content of calls or SMSs. The latitude and longitude in the dataset is at the mobile cell tower level, covering a range of 50 to 200 meters. All results are aggregated at the Singapore planning area level.

In this paper we propose a novel method based on inactivity to predict home location of a user using his mobile phone log records. We combined information from an open source (www.streetdirectory.com) to improve accuracy. Comparison results show that our method gives better accuracy than existing methods. To predict the work place we exploited the fact that a person goes to his/her work place on most of the weekdays and rarely on the weekends. Prediction of home and work-place can be used in urban planning. For example, home and work place distribution can be used to compute how **balanced** a planning area of a city is. A poorly balanced planning area will have biased home and work place distributions. Furthermore, distance travelled between home and work place can be used in urban planning.

II. Related Work

A. Prediction of home and work place

In [1] two cells with the highest regularity (i.e., number of days) are chosen as candidates for home and work place. The authors argued that home and work place can be distinguished by the standard deviation of start times of calls. It is assumed that usually one spends less time at work place than at home. So, it is expected that among the two candidates, the one with higher standard deviation is the home and the one with smaller standard deviation is the work place.

In SeMiTri [2], first of all, anchor points or stops (and moves) are determined from the raw GPS data. Next, each stop is mapped to a POI using a Hidden Markov Model (HMM). Open source data is used to determine the initial probabilities for each state in the HMM and the state transition matrix. So, in essence, SeMiTri uses correlation with open source data and HMM to determine home and work locations (and other POIs).

In the review paper [3], Shan Jiang et al. used frequency for a certain period to determine home and work place.

Possible duration for home is 9pm to 7am, and for work place is 12pm to 5pm.

B. Validation

Two types of validation are done: (a) select a sample of users who agree to disclose their home and work location for the study, and (b) use statistics from open source and compare them to statistics obtained from mobile phone data. It is difficult to obtain work place statistics from open source.

In [3] Shan Jiang et al. used the statistics obtained from the Massachusetts travel survey data to compare with their statistics.

In [1] the distribution of homes calculated by the model was compared to the data from the Estonian population register.

In [4] authors used travel survey data from Paris (23429 weekdays of people) and Chicago (23764 weekdays of people) (http://www.cmap.illinois.gov/travel-tracker-survey).

Nathan Eagle et al [5] used a small sample of 100 users (from MIT) over the course of nine months to validate the outcome of their research on sensing complex social systems from mobile network data.

The proposed methods in this paper are validated using both approaches. We also show how to validate work place.

C. Preprocessing

Preprocessing the mobile network data is an essential step to extract meaningful knowledge. It helps to focus on those users relevant for creating a useful model. Thresholds are used to remove users with very few or too many mobile network data, in other words outliers are filtered out. Another important preprocessing step is removing oscillation between towers.

In [1] authors argue that users with too few calls trying to predict home and work place is more of a speculation. Users having too many calls can be using some technical device such as GSM network, or it may be because of some organized call procedure (service centre, etc.).

In [3] after identifying *stay* regions, authors removed intermediate record (i.e., pass-by).In [2] and [6] also these steps are performed to preprocess the data.

Cell tower oscillation resolution becomes an essential preprocessing step because GSM network faces a critical problem of Cell oscillation where user is assigned different Cell IDs even when the user is stationary. It happens mainly because of two reasons: (a) load balancing effect, or (b) physically disruptive conditions (bodies of water, landscape height) [7].

III. Methods

Figure 1 shows a block diagram for home prediction and validation. A similar procedure is used for work place.

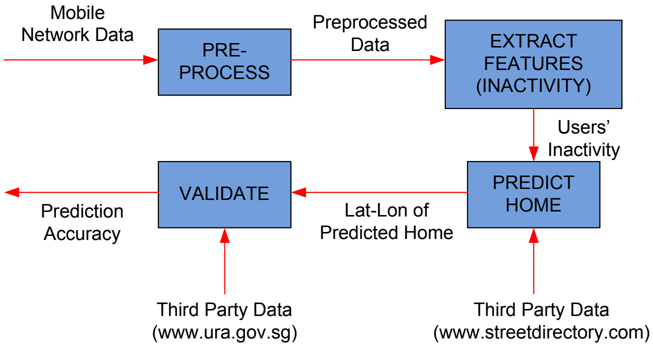
Mobile Network Data
PRE-PROCESS
Preprocessed Data
EXTRACT FEATURES (INACTIVITY)
Users' Inactivity
PREDICT HOME
Lat-Lon of Predicted Home
VALIDATE
Prediction Accuracy
Third Party Data (www.ura.gov.sg)
Third Party Data (www.streetdirectory.com)

Figure 1. Block diagram for home prediction method

A. Preprocessing

Removing Cell Tower Oscillation: We used a simple procedure to remove cell tower oscillation. If two consecutive records have the same time stamp or their time stamps are extremely close, we replace the tower having lower frequency with the tower having higher frequency.

Determining in transit Records: We used two thresholds (velocity and distance) to determine whether a mobile phone record is *in transit*. Velocity is measured by dividing distance between two consecutive records (i.e., two cell towers) with the time difference. If between two consecutive records, distance and velocity thresholds are greater than their respective thresholds, then the second record is *in transit*.

B. Method to Predict Home

We use inactivity to predict home location. Inactivity is defined as *no activity for more than threshold time except for location update event*. *InactivityThreshold* is set to five hours in order to model the sleeping hours. This works well for shift workers as well because it does not require the inactivity hours to be during night time. A location update event is an automated event initiated by the cellular provider in order to track the whereabouts of the mobile phone user. For every anonymized user we compute the total number of inactivities for each tower. The tower with the highest number of inactivity is predicted to be the home of the anonymized user. The algorithm to compute number of inactivities is given below.

We introduced a confidence measure to guard against imbalance in $inactivityCount$ between users with varying regularity. Regularity of a tower is defined as number of days this tower is used by the user.

$$confidence = \left[\frac{inactivityCount}{regularity}\right]^{1/3} \qquad (1)$$

If the difference between $regularity$ and $inactivityCount$ is high, the confidence in home prediction is low. Otherwise,

Algorithm 1 countNumInactivities algorithm

```
procedure COUNTNUMINACTIVITIES(records)
    for each anonymized ID $u$ do
        $inactivityCount = 0$
        for each record $r$ do
            $startTime = r.time$
            if $i^{th}$ tower = $(i+1)^{th}$ tower then
                if $(i+1)^{th}$ NOT locUpdEvent then
                    $startTime = (r+1).time$
                end if
                if $timeDiff(i, i+1) > inactThre$ then
                    Increment $inactivityCount$
                end if
            else
                if $timeDiff(i, i+1) > inactThre$ then
                    Increment $inactivityCount$
                    $startTime = (r+1).time$
                end if
            end if
        end for
    end for
end procedure
```

the confidence is high. The value of "1/3" is empirically determined.

Using countNumberOfInactivities() we select the location of the tower with the highest number of inactivities as home. Suppose, a user's home is near a non-residential area and the tower with the highest number of inactivities belongs to a non-residential area. Here we must understand that some users live in non-residential area for certain reasons (e.g., manufacturing workers). So, we cannot simply pick up the tower with the highest number of inactivities *among all residential towers* for a user as his home. We have to create a criterion for the trade-off. We sort the towers in descending order of their $inactivityCount$ for a user. Consider the top ranked residential tower (denote it i^{th} tower: all towers 0 to $i-1$ are non-residential). Check if the following condition satisfies for the i^{th} tower.

$$\frac{i^{th} \text{ tower } inactivityCount}{max(inactivityCount)} \leq \frac{1}{\beta} \tag{2}$$

Scenario 1: If the tower with the maximum $inactivityCount$ is of residential category, it is selected as home.
Scenario 2: Otherwise, if the above condition is satisfied, the top ranked tower is selected as the home. However, if the above condition is not satisfied, the i^{th} tower is selected as the home.

In the experimental section we discuss how to set β empirically. The above criterion ensures that if the difference between the top ranking residential tower and top ranking tower is high, the user may be living in a non-residential area.

C. Method to Predict Work-Place

Prediction of work-place of an anonymized user is more difficult than prediction of her home. The working hours differ significantly for different users. This is definitely true for shift workers. Even otherwise, those who have typical working hours of 9am to 5pm, they may come to work place and return home at different times. To add to this complexity, there are weekdays when a use goes to his work place and weekends when he does not. The proposed method is based on the fact that a user goes to his work place on weekdays regularly and rarely goes on weekends. We multiply this ratio by duration during weekday from 2pm to 5pm. By doing so, we give more importance to towers which have high duration during 2pm to 5pm on weekdays.

$$w = \frac{regularityWeekdays + k}{regularityWeekends + k} * durationWeekday \tag{3}$$

where $regularityWeekdays$ is the number of weekdays a tower is used by a user, $regularityWeekends$ is the number of weekends a tower is used by a user, k is a constant used to avoid division by zero condition, and $durationWeekday$ is the duration during weekdays from 2pm to 5pm.

IV. Experimental Results and Evaluation

A. Data

Large Network Data without Ground Truth: In this study we use mobile network data of 3,875,254 anonymized users. These users are locals of Singapore.There are around eight billion records over three months (May - July, 2013). A mobile phone record includes the following fields among others: anonymized ID, time stamp, latitude and longitude of cell tower. We do not know the ground truth (i.e., home and work place) of users.

Sample Data with Ground Truth: A sample of 4515 anonymized users agreed to disclose their home locations.

B. Methods

Validation Procedure: We used publicly available data: StreetDirectory.com and Urban Redevelopment Authority (www.ura.gov.sg) (URA), for validation. StreetDirectory.com maps geographical locations of Singapore (latitude, longitude) to semantic categories such as HDB blocks, condominium, semidetached house, shopping mall, shop houses, university, military, civil defence camp. URA provides planning area-wise distribution of population in Singapore. The following procedure is used for validation of home prediction.

Validation Method for Home Predictions

Algorithm 2 Validation algorithm

```
procedure VALIDATE(records, 3rd party statistics)
    for each anonymized ID do
        Predict home location (latitude and longitude)
        if maximum inactivity tower not residential then
            Find the highest inactivity residential tower
        end if
        Find planning area from StreetDirectory.com
    end for
    Compute number of users for each planning area
    Compare this statistics with URA statistics
    Compute sum(difference in percentages) for all areas
    Compute correlation
end procedure
```

Comparison: We compared the proposed method with the method in [1].It is based on the idea that typically working hours do not spread over the entire 24 hours rather they are limited to a certain eight hour period in a day. So, if we take standard deviation of the time stamps of each phone event for each tower, the tower corresponding to the work place will have lower standard deviation than the home tower. It requires two thresholds - for average and standard deviation of time stamp. We empirically set these two parameters so that the prediction accuracy is maximized.The best threshold for average time stamp is 19:00 and threshold for standard deviation of time stamp is 4.0.

We also compared our method with [3].

C. Results and Validation

1) Home Prediction and Validation for Large Data without Ground Truth: The results are summarized as follows.

We predicted home locations for 3,875,254 anonymized users. Home locations were grouped by their planning areas. Then, the predicted distribution was compared to the distribution of URA. The 2012 URA statistics shows distribution of 3,818,200 people in Singapore. Table I shows the details of the distributions. In order to compare the two statistics we used percentages for each planning area. We computed absolute sum of error (i.e., difference) between the two distributions using the formula given below.

$$AbsSumError = \sum_{i=1}^{nd} |P_i - U_i| \qquad (4)$$

where nd is number of planning areas (including the junk class "others"), P_i is the predicted statistic value (in percentage) for i^{th} planning area and U_i is the URA statistic value (in percentage) for i^{th} planning area.

AbsSumError is reduced slightly (about 0.5%) by the proposed method vis-a-vis Ahas-etal. But when we combine with open source information (www.streetdirectory.com), the error is reduced by (approx.) 8% for both methods (see Table I).

Correlation is given as follows [8]:

$$r = \frac{1}{(nd-1)} \sum_{i=1}^{nd} \left(\frac{P_i - PM}{sd_P} \right) \left(\frac{U_i - UM}{sd_U} \right) \qquad (5)$$

where r is Pearson's correlation coefficient, PM is mean predicted statistic value, UM is mean URA statistic value, sd_P is standard deviation of predicted statistic values and sd_U is standard deviation of URA statistic values. Arguably correlation captures difference between two distributions better than absolute sum of errors.

Correlation coefficient of both methods are very close. But it improves significantly when we combine with open source information (see Table I).

2) Home Prediction and Validation for Sample Data with Ground Truth: We run the proposed method over a sample of 4515 anonymized users who agreed to declare their home locations. Our predictions were correct (within 3km of declared home location) for 85.5% users. However the distance from the declared and predicted home locations for the remaining 665 anonymized users is greater than 3km. We conducted further investigations for these 665 users. For each person, we computed the average distance between home (predicted or declared) and visited places during night time (9pm-7am). Say, during one night a user visits $(P_1, P_2, ..., P_h)$ places. Then, average distance to predicted home is calculated as follows:

$$distance = \frac{1}{h} \sum_{i=1}^{h} dist(Home_{predicted}, P_i) \qquad (6)$$

Similarly, average distance to declared home is calculated as follows:

$$distance = \frac{1}{h} \sum_{i=1}^{h} dist(Home_{declared}, P_i) \qquad (7)$$

It was found that for more than two-thirds of 665 anonymized users distance from night location to predicted home is less than distance to declared home (see Figure 2). This case may typically occur for residents such as students who would stay on campus during week days and stay home during week ends.

Our analysis showed that, particularly for a city like Singapore, because of high population density, chances of users staying close to non-residential area is high. We applied Equation 2 to improve the accuracy of prediction for such users. β is set to $\frac{1}{3}$. Accuracy improved from 85.5% to 88%.

3) Work Place Prediction and Validation: We predicted work place using equation 3. Unfortunately validation of work place is not easy. We wanted planning area-wise distribution of all job sectors. But there is no such information except for one sector, i.e., manufacturing. Planning area-wise

Table I
PLANNING AREA-WISE DISTRIBUTION OF SINGAPORE: GROUND TRUTH, PREDICTION BY PROPOSED METHOD AND AHAS-ETAL-2010

	Ground Truth (URA)		Proposed Method				Ahas-etal-2010 Method			
			w/o open source data		with open source data		w/o open source data		with open source data	
Planning Area	Total	%	%	Diff	%	Diff	%	Diff	%	Diff
Total	3,818212		100% = 3,875,254		100% = 3,875,254		100% = 3,875,254		100% = 3,875,254	
Ang Mo Kio	178901	4.685	3.82	0.862	3.94	0.747	3.46	1.222	3.95	0.736
Bedok	295244	7.731	5.89	1.845	6.15	1.585	5.18	2.552	6.26	1.467
Bishan	92791	2.428	1.86	0.566	1.98	0.445	1.68	0.749	1.97	0.456
Bukit Batok	142651	3.735	2.93	0.807	3.02	0.712	2.63	1.100	3.08	0.653
Bukit Merah	157214	4.117	4.02	0.093	4.49	0.369	4.31	0.191	4.44	0.320
Bukit Panjang	131000	3.431	2.25	1.183	2.30	1.131	1.78	1.651	2.27	1.157
Bukit Timah	71918	1.883	2.11	0.227	2.53	0.645	2.00	0.119	2.47	0.592
Changi	2501	0.065	1.05	0.981	0.17	0.103	1.38	1.312	0.21	0.147
Choa Chu Kang	174236	4.562	2.87	1.697	3.34	1.218	2.28	2.283	3.41	1.155
Clementi	91253	2.389	1.96	0.429	2.28	0.110	1.78	0.611	2.29	0.100
Downtown	3716	0.097	1.81	1.710	1.22	1.122	4.36	4.266	1.38	1.280
Geylang	119332	3.125	4.51	1.389	3.70	0.573	4.56	1.438	3.89	0.767
Hougang	217452	5.694	4.47	1.222	5.00	0.697	3.99	1.703	5.06	0.633
Jurong East	86591	2.265	2.71	0.447	2.84	0.578	2.77	0.503	2.82	0.556
Jurong West	271973	7.121	6.16	0.963	7.20	0.081	5.30	1.825	7.30	0.179
Kallang	102962	2.695	3.14	0.448	4.24	1.547	3.42	0.729	4.60	1.909
Mandai	2102	0.055	0.04	0.011	0.09	0.030	0.04	0.013	0.07	0.012
Marine Parade	48500	1.270	1.43	0.159	1.35	0.081	1.31	0.043	1.34	0.071
Newton	6511	0.170	0.32	0.150	0.28	0.105	0.40	0.231	0.28	0.110
Novena	47142	1.234	1.96	0.727	1.88	0.648	2.00	0.766	1.87	0.636
Outram	22020	0.576	0.87	0.291	0.92	0.339	1.18	0.605	0.93	0.352
Pasir Ris	135987	3.559	2.69	0.874	2.87	0.692	2.29	1.274	2.89	0.669
Punggol	74701	1.956	1.47	0.487	1.63	0.329	1.24	0.713	1.72	0.241
Queenstown	97802	2.561	3.22	0.654	3.05	0.486	4.01	1.449	3.04	0.476
River Valley	8606	0.225	0.66	0.434	0.30	0.072	0.64	0.416	0.29	0.064
Rochor	15234	0.398	1.73	1.334	1.83	1.428	1.95	1.550	2.30	1.898
Sembawang	73305	1.920	1.98	0.061	1.66	0.256	1.95	0.029	1.70	0.216
Sengkang	177865	4.657	3.17	1.489	3.21	1.446	2.74	1.919	3.39	1.262
Serangoon	123323	3.229	2.68	0.549	2.93	0.301	2.59	0.641	3.04	0.192
Singapore River	2301	0.060	0.38	0.320	0.40	0.341	0.49	0.433	0.49	0.429
Tampines	260000	6.809	5.40	1.411	5.23	1.582	4.96	1.847	5.20	1.610
Tanglin	17810	0.466	1.15	0.686	1.32	0.853	1.18	0.713	1.37	0.905
Toa Payoh	126221	3.305	2.63	0.680	2.92	0.383	2.46	0.844	2.98	0.320
Woodlands	247806	6.490	4.95	1.539	5.36	1.126	4.21	2.284	5.56	0.930
Yishun	187202	4.903	3.62	1.283	3.67	1.235	3.21	1.698	3.85	1.052
Others	4900	0.128	8.10	7.974	4.72	4.593	10.27	10.139	4.78	4.656
			Absolute Error = 35.983		Absolute Error = **27.5**		Absolute Error = 36.1		Absolute Error = 28.2	
			Corr Coeff = 0.68		Corr Coeff = **0.87**		Corr Coeff = 0.69		Corr Coeff = 0.86	
Without “Others”:			Corr Coeff = 0.93		Corr Coeff = **0.95**		Corr Coeff = 0.92		Corr Coeff = 0.93	

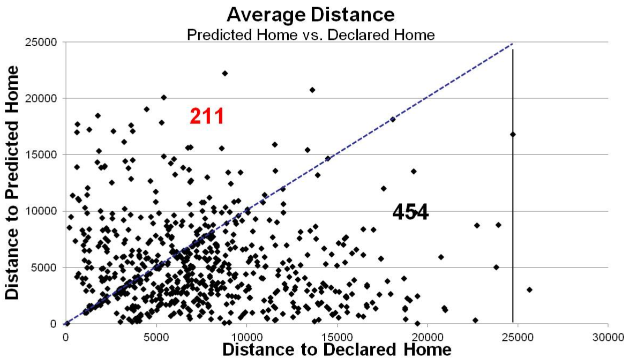

Figure 2. For more than two-third of 665 anonymized users, our predictions associate with night activities well

distribution of manufacturing sector for both ground truth (Source: Department of Statistics, DoS, Singapore) and prediction by our proposed method are compared. AbsSumError is 27.05% and correlation coefficient is 0.92.

D. Experiments on Scalability and Performance

It is essential to check scalability of our proposed method as the data may contain billions of records. In Figure 3 we show time taken for 20%, 40%, 60%, 80% and 100% records. Increase in time is **linear** with percentage of data.

We also tested performance of the proposed method for home prediction accuracy. In Figure 4 we show prediction accuracy for varying inactivity duration. We set inactivity duration to 4hr, 5hr and 6hr. It clearly shows performance is the best for (inactivity duration = 5hr) both for absolute

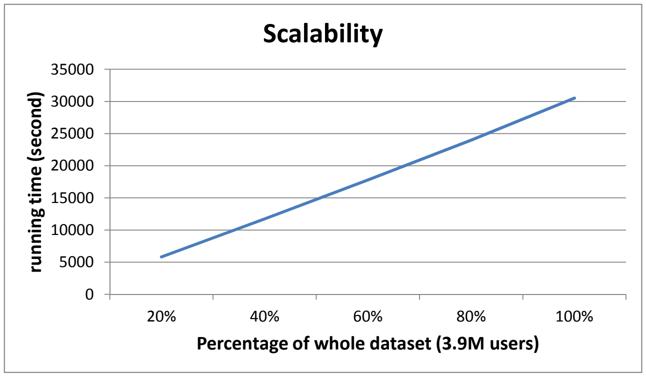

Figure 3. Test for scalability of the proposed method

error and correlation coefficient. So, in all our experiments we set inactivity duration to 5hr.

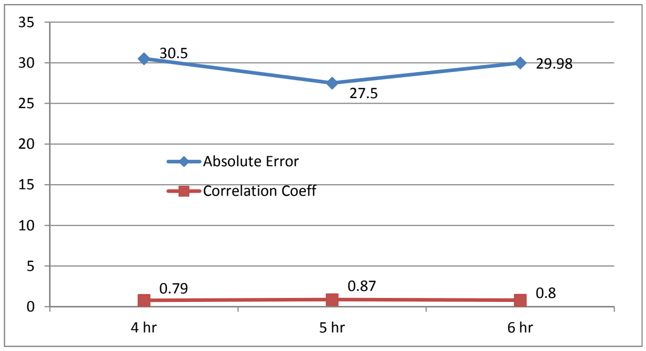

Figure 4. Performance (absolute error and correlation coefficient) of the proposed method for time = {4hr, 5hr, 6hr}

V. Discussion on Urban Planning

Prediction of home and work places can assist in urban planning.

- Home-Work Balance: Balance of a planning area P is defined as the ratio of people working in planning area P who also live in P and total working population living in P. A more "balanced" planning area will be able to absorb its own population working in that planning area itself. Using the prediction of the proposed method, a well balanced planning area is Tampines. This result is supported by the **World Habitat Award** by united nations[1]. According to this study Jurong Island, which is largely a manufacturing and trading planning area, has low balance. Modern urban planning strongly favors high population density with development, which translates to high balance.
- Travel Distance: A urban development planner will like to know distribution of distance travelled from home to work place. Using our system, the top five planning areas with the highest average trips from home to work are (in descending order): North-Eastern Islands, Lim Chu Kang, Simpang, Sungei Kadut and Western Water Catchment. The planning areas with the lowest average trips are (in ascending order): River Valley, Newton, Rochor, Museum, and Downtown Core. These results are very reasonable. For example, people of downtown core travel less distance to work, but people living in places like north-eastern islands have to travel a long distance to work.

[1] http://en.wikipedia.org/wiki/Tampines and http://www.hdb.gov.sg/fi10/fi10320p.nsf/w/AboutUsTown Tampines

References

[1] R. Ahas, S. Silm, O. Jrv, E. Saluveer, and M. Tiru, "Using mobile positioning data to model locations meaningful to users of mobile phones," *Journal of Urban Technology*, vol. 17, no. 1, pp. 3–27, 2010.

[2] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer, "Semitri: a framework for semantic annotation of heterogeneous trajectories," in *Proceedings of the 14th international conference on extending database technology*. ACM, pp. 259–270.

[3] S. Jiang, G. A. Fiore, Y. Yang, J. Ferreira Jr, E. Frazzoli, and M. C. Gonzalez, "A review of urban computing for mobile phone traces: current methods, challenges and opportunities," in *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. ACM.

[4] A. Wesolowski, N. Eagle, A. M. Noor, R. W. Snow, and C. O. Buckee, "The impact of biases in mobile phone ownership on estimates of human mobility," *Journal of The Royal Society Interface*, vol. 10, no. 81, 2013.

[5] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and ubiquitous computing*, vol. 10, no. 4, pp. 255–268, 2006.

[6] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang, "Collaborative location and activity recommendations with gps history data," in *Proceedings of the 19th international conference on World wide web (WWW '10)*. ACM, pp. 1029–1038.

[7] S. A. Shad and E. Chen, "Cell oscillation resolution in mobility profile building," *CoRR*, vol. abs/1206.5795, 2012.

[8] M. Pagano and K. Gauvreau, *Principles of Biostatistics*, 2nd ed. Duxbury Thomas Learning, 2000.