



AMDA: Anchor Mobility Data Analytic for Determining Home-Work Location from Mobile Positioning Data

A P Putra¹, W O Z Madjida², I A Setyadi², A R S Nugroho², A R MNSP Munaf²

¹ Directorate of Statistics Methodology, BPS - Statistics Indonesia - Jakarta, Indonesia

² Directorate of Statistical Information System, BPS - Statistics Indonesia - Jakarta, Indonesia

*Corresponding author's e-mail: amanda.putra@bps.go.id, zuhayeni@bps.go.id, adityasetyadi@bps.go.id, sinung@bps.go.id, alfa@bps.go.id

Abstract. In conducting a mobility analysis using Mobile Positioning Data, the most critical step is to define each customer's usual environment. The initial concept of mobility used is the movement that occurs from and to every usual environment, so errors in determining the usual environment will cause incorrect mobility statistics. Therefore, Anchor Mobility Data Analytic (AMDA) is proposed for Home-Work Location Determination from Mobile Positioning Data. This algorithm uses clockwise reversal to make it easier to classify someone in their usual environment. Unfortunately, only about 80% of the raw data can be used to establish usual environments. The remaining 20% do not have sufficient data history. This study found that the accuracy of AMDA in determining monthly home location was 98.8% at the provincial level and 88.7% at the regency level. As for the determination of monthly work locations, 98.9% at the provincial level and 70.4% at the regency level.

1. Introduction

The number of studies about human mobility increased around 2008 and is still multiplying [1]. Human mobility is becoming increasingly complex and significant because it affects various aspects of human life, such as the spread of disease and viruses (e.g., COVID-19) [2] [3] [4], people's behaviour after a disaster [5] [6], commuting behaviour [7] [8], as well as transport and traffic volume [9] [10]. It also influences some decisions in the tourism sector [11] and other urban mobility planning [12][13]. These various aspects are based on daily movement patterns such as where a person lives, where they do their daily activities (work, school, shopping, and other activities.), or to a new place to visit. The place where you live or where you regularly visit is called usual environment. Usual environment is the geographical area in which an individual carries out his/her routine of life [14]. The usual environment of an individual includes usual residence of the household to which he/she belongs (place of home), his/her place of work or study, and any other place that he/she visits regularly and frequently, even when this place is located far away from his/her place of usual residence. Home location is an important place that becomes the first step for further human mobility studies [15], likewise with work, which is the basis for determining the commuting, transportation, and traffic planning analysis.

Assessment and analysis of individual locations in the current mobile era require new methods and approaches. Traditional surveys and population registers are not flexible on everyday mobility. Currently, Mobile Positioning Data is increasingly considered a new and exciting source of



information to study the spatial dynamics of human society. This is also supported by around 66.9% of the world's population using mobile phones. Some of the uses of mobile positioning include determination of origin-destination and human mobility [16][17][18]. Other studies have also developed various approaches to detect the usual environment from Mobile Phone Data, especially the home and work location [19]-[22]. Most of the proposed home and work detection algorithm uses Call Detail Record (CDR), a billing log recorded for called activities. However, CDRs have significant limitations as a source of location information. CDRs records are sparse because they are generated only when there is a voice call or text message. In the current era, where most mobile users use the internet more, even when making calls, the number of records generated on the CDR is insignificant compared with other mobile activities.

We proposed another mobile positioning data source such as Location-Based Advertising (LBA) to infer each individual's home and workplace. The sources will refine the data from CDR to describe the mobility patterns of mobile users better. We modified the algorithm in [20] by reducing some aspects of the data bias. The first modification is adding a day filter, i.e., removing weekends in processing. We assume that activities carried out on weekdays can provide a clear separation between home and work compared to weekends. This is because, on weekends, users may spend more time at home or tourist attractions to affect the calculation on anchor modelling. In addition, we also clustered locations that are considered close to the primary candidate for minimizing the exact location being considered a different anchor just because users are detected by different Base Transceiver Station (BTS). Then, the results of our proposed home and work detection algorithm will be evaluated at the individual level as ground truth by comparing the locations of the algorithm results with the actual home and office locations of each individual.

The organization of this paper is as follows: Section 2 explains more detail about mobile positioning data sources that we used, our proposed algorithm, and how we validate it. Then, experiment results are reported in Section 3. Finally, we give concluding remarks in Section 4.

2. Data and methodology

2.1.1 Passive mobile positioning. One of the most promising Information and Communication Technologies (ICT) data sources for assessing human movement and mobility is mobile positioning data. In today's environment, nearly everyone on the earth uses a cell phone in their daily activities. As a result, every action leaves a digital footprint, which can grow rather large over time.

Active mobile positioning data is mobile tracing data in which the phone's location is determined (asked) via a specific radio wave query. A particular environment and permission from the phone owner are necessary to inquire about the whereabouts of particular phones. On the other hand, mobile operators' passive mobile positioning data is automatically stored in the log files (billing memory; hand-over between network cells, Home Location Register, etc.), known as a call detail record.

Call Detail Record (CDR) is collected by the MNO regularly for processing into usage, capacity, performance, and diagnostic reports. However, in this study, we would like to gain helpful insight from this data. CDR is an automatically generated log that contains active usage of the mobile phone in the network, such as incoming and outgoing calls, SMS, GPRS, etc. The recorded transaction data is also accompanied by information on the time of the activity, followed by customer identification in the form of a mobile phone number. Location identification is established from records of the BTS location of the cellular network provider in the form of position coordinates. Compared with active positioning data, the spatial accuracy of passive mobile positioning data is much lower, and the spatial interval is usually irregular and with more extended time gaps.

Location information generated from LBS (Location Based Services) data by MNO usually contains three different types of data sources, such as: Charging Data (CHG); Location-Based Advertising (LBA); and Unified Policy and Charging Controller (UPCC) (see Table 1).

**Table 1.** Three basic types of data sources generated from LBS.

Source Type	Description	Alias
CHG	Billing domain log which stores successful charging transaction record such calls, messaging, etc.	calls, sms, and mms logs
Source Type	Description	Alias
LBA	The technology is used to pinpoint consumers location and provide location-specific advertisements on their mobile devices.	signalling data
UPCC	It provides policy, service, subscription, quota, and bearer resource management functions, as well as admission control for internet data usage.	internet data usage

2.2 Data processing

The digital footprint left by users from mobile positioning data is incredibly sensitive. However, it is also highly essential since it offers new ways to quantify and track the spatiotemporal activities of the population. On the other hand, exploiting each individual's movement and mobility to get insight may jeopardize freedom of movement and private rights. As a result, privacy concerns must be addressed by utilizing appropriate privacy protection mechanisms, such as anonymizing the subscriber's phone number. This technique is usually called the anonymization technique and is applied in the early stages of the data processing process. It is applied following telecommunications laws and regulations, which are not allowed to share personal data with other parties without the person's consent.

Mobile positioning data are still open to being compromised through various problems: human error, whether malicious or unintentional; transfer errors, including unintended alterations or compromise from one to another; bugs, viruses/malware, hacks, and other cyber threats; compromised hardware, such as a device or disk crash; physical compromise to devices or subscriber's phone. Therefore, it is necessary to do quality control and essential data cleansing before data processing. Data cleansing is an essential part of the data processing process since it improves data quality and, as a result, increases overall productivity. In addition, this phase does eliminate any obsolete or erroneous data, leaving only specific highest-quality data.

Reverse geocoding generates street addresses, descriptive places, or administrative regions from latitude and longitude data. The ability to reverse geocode coordinates on mobile location-based data would benefit people's mobility analytics. When converted to addresses or administrative areas, it will be easier to discover movement and interaction supplied by GPS coordinates. We may extract their point locations by overlaying their generated coordinate location on the data with the BPS village master spatial layer, then matching it with administrative areas, which are utilized to identify individuals (see Figure 1).

datetime	msisdn	latitude	longitude
13/1/2020 06:35	628xxxx	-6.28247	106.8734
13/1/2020 08:14	628xxxx	2.18461	117.4982
13/1/2020 19:04	628xxxx	-0.40122	116.5368
13/1/2020 07:59	628xxxx	2.23036	117.496
13/1/2020 08:02	628xxxx	2.23036	117.496
13/1/2020 07:55	628xxxx	2.23036	117.496
13/1/2020 08:06	628xxxx	2.23036	117.496

Figure 1. Sample generated transaction records.



2.3 The baby steps

2.3.1 Anchor model

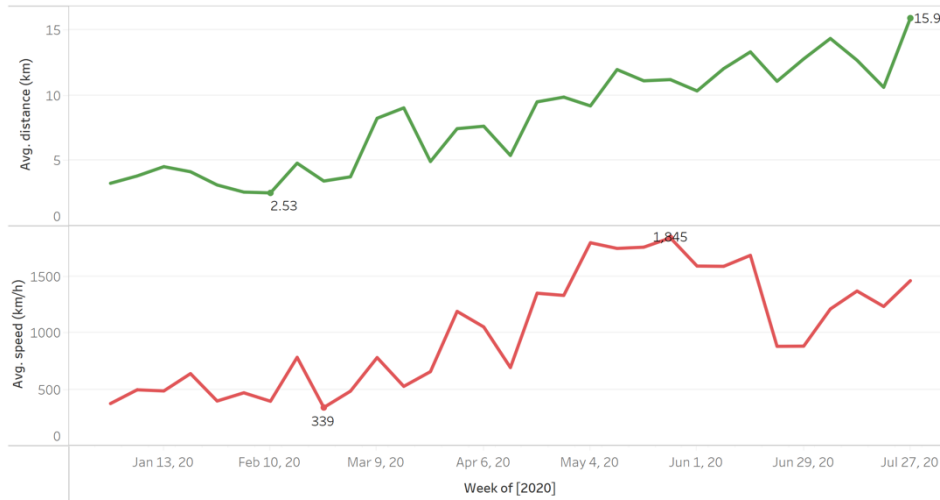


Figure 2. Weekly average speed and distance, travel included.

The first stage of forming the anchor model is to remove the high-speed location transfer data from raw data to raw master data by calculating the sequence distance of each MSISDN location by applying the Haversine distance between them, then calculating the speed between locations. The raw master drops data at what is considered an unusual speed, with a maximum threshold of 10 km/h (assuming the average time it takes a person to run). In Figure 3, we see that the site's movement speed is considered constant in the range of 0.54km/h to 1.51km/h, with a distance of 0.39km to 1.48km.

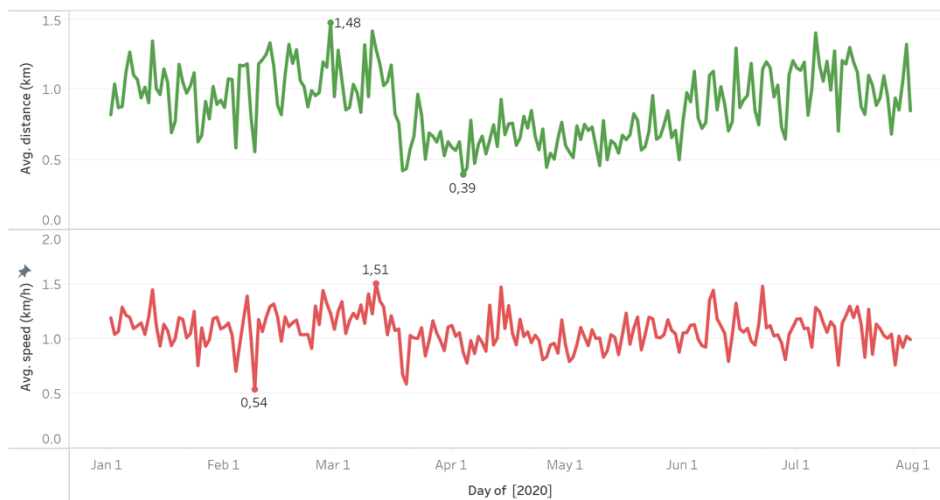


Figure 3. Daily average speed and distance which approximately stays.

Algorithm for home detection is using AMDA-Home which is modification of Anchor model [20]. The stages are explained as follows:

1. For every subscriber, filter the first daily recorded events in every cell-id that occur on weekdays.



2. Converts event time to clockwise reversible form: transform the time stamp that has 24 hours in raw data as we present in figure 4. The transformation is conducted using the following algorithm:
 - a. For every timestamp, check whether hour > 12.
 - b. If 'yes' (hour > 12), then timestamp is transformed using a formula: hour – 12.
 - c. If 'no', timestamp is not transformed.
 - d. Next step, for every subscriber calculate the frequency of being in a certain LAU (in this case LAU2 regency).

Original	0	1	2	3	4	5	6	7	8	9	10	11
	23	22	21	20	19	18	17	16	15	14	13	12
Converted	0	1	2	3	4	5	6	7	8	9	10	11

Figure 4. Clockwise reversible.

3. Calculate the frequency of initial events per day per cell-id (anchor).
4. Calculates the statistical mean and standard deviation for the time the event occurred, the number of days, and the number of occurrences of the event. This calculation is performed for each anchor.

2.3.2 *Usual Environment Candidate.* After determining the anchors, the next step is to determine the usual environment candidates. The stages are as follows:

1. Calculate the distance between the anchor and the main candidate.
2. Filter anchors with the number of days appearing that is more than or equal to 5. Eliminate anchors that are close to the main candidate (below 500m).
3. Segregation of events only on weekdays, as well as transformation of time variables into 12-hour format.
4. Calculation of the frequency of the daily initial location of each location point (anchor) followed by the statistics of the average and standard deviation of each hour of the incident, and made the main candidate.
5. Calculation of the distance between the selected locations and the main candidate locations and then eliminating adjacent locations that are below 500 m, followed by sorting the days of occurrence above or equal to 5 days.

In Anchor 1, the data is filtered only to use the weekday's data. The hour's feature also transforms into a clockwise reversible format, as shown in Figure 4. After that, the number of records (N_event), the number of unique days (N_date), as well as the average hour of occurrence (AVG_hour) followed by the standard deviation of the hour (SD_hour) are calculated for every subscriber per month. All of these features will be used to determine the usual environment. In specific, the standard deviation will then be used as the final filter for a usual environment when the data have the same number of days, number of data lines, and average hours. The results of data processing formed at the anchor 1 stage are presented in figure 5.

msisdn	year	month	latitude	longitude	idkab	N_event	N_date	AVG_hour	SD_hour
628xxxx	2020	1	-6.9201	107.5607	32 77	113	22	1.05	2.07
628xxxx	2020	1	-6.922	107.5559	32 77	46	22	3.83	3.03
628xxxx	2020	1	-6.9201	107.5607	32 77	155	19	3.6	3.55
628xxxx	2020	1	-6.922	107.5559	32 77	67	18	4.31	4.05
628xxxx	2020	1	-7.0282	107.5228	32 04	105	12	8.89	0.8

Figure 5. Anchor 1 aggregation result.



In Anchor 2, we will check the number of occurrences of a location with a minimum approach of appearing for at least five days monthly. Furthermore, this stage will examine each of these potential locations by grouping/clustering each existing location. This grouping is done by calculating the distance between locations (using the Haversine distance), then grouping the locations with a distance radius of 0.5 km. Then, the location considered close to the primary candidate of the usual environment (illustration in Figure 6) is omitted from the candidate of the usual environment.



Figure 6. Anchor clustering.

The usual environment is defined into two locations, which are home and office. The anchor point with an average hour is less than seven ($AVG_hour < seven$) will be classified as the candidate of home location, whether the other anchors with an average hour greater than seven will be classified as the candidate of work location. An example of the process of determining this classification is illustrated in Figure 7.

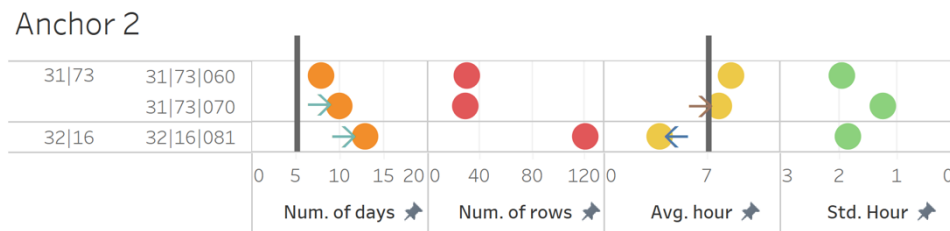


Figure 7. Usual Environment Selection Criteria.

After each anchor point has been grouped into a prospective home or work, select the strongest candidate for each group to predict the home and work. The criteria for determining the prediction are: select the highest number of days of occurrence (N_date), if they are the same, then choose the location with the most data rows (N_event), if it is still the same, use the highest average hour (AVG_hour), and if it is still the same, then select the lowest standard deviation the hours (SD_hour). Finally, obtain the prediction of the usual environment, as shown in Figure 8.

msisdn	year	month	latitude	longitude	idkab	N_event	N_date	AVG_hour	SD_hour	Label
628xxxx	2020	1	-6.25008	106.9067	31 72	39	22	1.07	1.96	Home
628xxxx	2020	1	-7.0282	107.5228	32 04	105	12	8.89	0.8	Work

Figure 8. Prediction of usual environment.

2.4 Validation

We check the accuracy of the resulting algorithm by testing the model on 992 volunteers who participated in the Metropolitan Statistical Area research in the Cekungan Bandung area in 2019. First,



the volunteers were asked the location of the usual environment through a survey, then the results of this survey were compared with the usual environment from MPD processing. It should be noted that some volunteers incorrectly determined their usual environment in answering the survey. The reason is that they live on the border and do not know the exact boundaries of their territory, nor do they know the time limit of staying in an area to be visited called the normal environment.

3. Result

3.1 Anchor Model Processing

Table 2. The amount of data formed at each stage.

Example Period	Raw	Master Raw	Anchor 1	Anchor 2
January (% to raw data)	100,00%	99,93%	98,01%	81,49%
February (% to raw data)	100,00%	100,00%	96,81%	79,52%

For example, we used two months of data to see what happened at the AMDA formation stage. We can see from Table 2 a significant data reduction (in the range of 20% of raw data) at the stage of forming Anchor 2. This happens because in Anchor 2, we choose subscribers that can form the usual environment, where one of the main requirements is to have sufficient historical data. In other words, we omitted phone numbers that we thought were disposable phone numbers. *Usual Environment Accuracy*

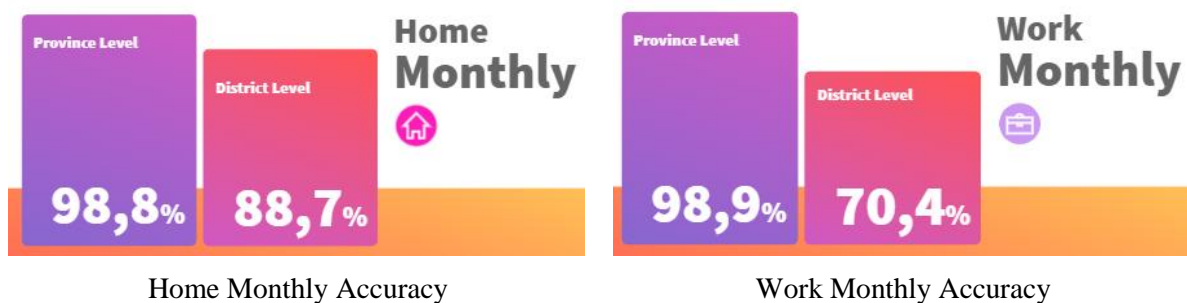


Figure 9. AMDA Accuracy in Province and District Level.

We validate the suitability of the anchor point model in determining the usual environment by testing it on some volunteers whose locations were known. Based on the analysis, we know that the anchor point model that has been developed can detect the location of the usual environment with an accuracy rate of 98.8% for the home location at the provincial level and 88.7% at the district/city level. Furthermore, the office's location can be predicted with an accuracy of 98.9% at the provincial level and 70.4% at the regency level.

In the validation stage, we found several sources of error. Among them, volunteers who live on Bandung City and Bandung Regency border tend to say they live in Bandung City, while their digital footprint proves that they live in Bandung Regency. Another finding is that if the move is new, volunteers who move, based on historical data, will be recorded as having a typical environment at the old address. In addition, there is an error in guessing the location of their home and workplace for night shift workers because, naturally, they have different activity patterns.

4. Discussions and Conclusions

In this study, the validation stage was carried out only by comparing the volunteer's acknowledgment of the usual environment and the digital footprint of the MPD. For future research, it is necessary to consider applications that use active mobile positioning, such as verification to Google timeline or travel diaries applications. The evaluation of the algorithm and its development should be assessed at



the level of accuracy in a more specific domain, for example at the sub-district level or at the village level.

If you want to replicate in other operators, it is necessary to consider the availability and completeness of the data before applying the AMDA algorithm.

Acknowledgments

The study of the use of mobile positioning data as data source of official statistics was carried out by BPS together with Telkomsel, a cellular operator which is a subsidiary of Telkom, a State-Owned Enterprise.

References

- [1] Barbosa, H., Barthelemy, M., Ghoshal, G., James, C.R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J.J., Simini, F. and Tomasini, M., 2018. Human mobility: Models and applications. *Physics Reports*, 734, pp.1-74.
- [2] Lai, S., Farnham, A., Ruktanonchai, N.W. and Tatem, A.J., 2019. Measuring mobility, disease connectivity and individual risk: a review of using mobile phone data and mHealth for travel medicine. *Journal of travel medicine*, 26(3), p.taz019.
- [3] Kraemer, M.U., Yang, C.H., Gutierrez, B., Wu, C.H., Klein, B., Pigott, D.M., Du Plessis, L., Faria, N.R., Li, R., Hanage, W.P. and Brownstein, J.S., 2020. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*, 368(6490), pp.493-497.
- [4] Oliver, N., Lepri, B., Sterly, H., Lambiotte, R., Deletaille, S., De Nadai, M., Letouzé, E., Salah, A.A., Benjamins, R., Cattuto, C. and Colizza, V., 2020. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle.
- [5] Wang, Y. and Taylor, J.E., 2018. Coupling sentiment and human mobility in natural disasters: a Twitter-based study of the 2014 South Napa Earthquake. *Natural hazards*, 92(2), pp.907-925.
- [6] Song, X., Zhang, Q., Sekimoto, Y., Shibasaki, R., Yuan, N.J. and Xie, X., 2016. Prediction and simulation of human mobility following natural disasters. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(2), pp.1-23.
- [7] Kung, K.S., Greco, K., Sobolevsky, S. and Ratti, C., 2014. Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one*, 9(6), p.e96180.
- [8] Stigell, E., 2011. Assessment of active commuting behaviour: walking and bicycling in Greater Stockholm (Doctoral dissertation, Örebro universitet).
- [9] Di Lorenzo, G., Sbodio, M., Calabrese, F., Berlingerio, M., Pinelli, F. and Nair, R., 2015. All aboard: Visual exploration of cellphone mobility data to optimise public transport. *IEEE transactions on visualization and computer graphics*, 22(2), pp.1036-1050.
- [10] Phithakkitnukoon, S., Sukhvibul, T., Demissie, M., Smoreda, Z., Natwichai, J. and Bento, C., 2017. Inferring social influence in transport mode choice using mobile phone data. *EPJ Data Science*, 6, pp.1-29.
- [11] Raun, J. and Ahas, R., 2016, November. Defining usual environment with mobile tracking data. In 14th Global forum on tourism statistics (pp. 23-25).
- [12] McNeill, G., Bright, J. and Hale, S.A., 2017. Estimating local commuting patterns from geolocated Twitter data. *EPJ Data Science*, 6, pp.1-16.
- [13] Osorio-Arjona, J. and García-Palomares, J.C., 2019. Social media and urban mobility: Using twitter to calculate home-work travel matrices. *Cities*, 89, pp.268-280.
- [14] IRTS. 2008. International Recommendations for Tourism Statistics 2008. New York: UNWTO
- [15] Bojic, I., Massaro, E., Belyi, A., Sobolevsky, S. and Ratti, C., 2015, December. Choosing the right home location definition method for the given dataset. In *International Conference on Social Informatics* (pp. 194-208). Springer, Cham.
- [16] Alexander, L., Jiang, S., Murga, M. and González, M.C., 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation research part c: emerging technologies*, 58, pp.240-250.
- [17] Demissie, M.G., Antunes, F., Bento, C., Phithakkitnukoon, S. and Sukhvibul, T., 2016, June. Inferring origin-destination flows using mobile phone data: A case study of Senegal. In 2016



- 13th International conference on electrical engineering/electronics, computer, telecommunications and information technology (ECTI-CON) (pp. 1-6). IEEE.
- [18] Yang, X., Fang, Z., Xu, Y., Shaw, S.L., Zhao, Z., Yin, L., Zhang, T. and Lin, Y., 2016. Understanding spatiotemporal patterns of human convergence and divergence using mobile phone location data. *ISPRS International Journal of Geo-Information*, 5(10), p.177.
- [19] Ahas, R., Silm, S., Saluveer, E. and Järvi, O., 2009. Modelling home and work locations of populations using passive mobile positioning data. In *Location based services and TeleCartography II* (pp. 301-315). Springer, Berlin, Heidelberg.
- [20] Ahas, R., Silm, S., Järvi, 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.
- [21] Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J. and Varshavsky, A., 2011, June. Identifying important places in people's lives from cellular network data. In *International conference on pervasive computing* (pp. 133-151). Springer, Berlin, Heidelberg.
- [22] Kung, K.S., Greco, K., Sobolevsky, S. and Ratti, C., 2014. Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one*, 9(6), p.e96180.