

Where did you come from? Where did you go?

Robust policy relevant evidence from mobile network big data

Danaja Maldeniya, Amal Kumarage, Sriganesh Lokanathan,
Gabriel Kreindler, Kaushalya Madhawa

March 2015



LIRNEasia is a pro-poor, pro-market think tank whose mission is *Catalyzing policy change through research to improve people's lives in the emerging Asia Pacific by facilitating their use of hard and soft infrastructures through the use of knowledge, information and technology.*

Contact: 12 Balcombe Place, Colombo 00800, Sri Lanka. +94 11 267 1160.

info@lirneasia.net

www.lirneasia.net



IDRC | CRDI

International Development Research Centre
Centre de recherches pour le développement international



Canada

This work was carried out with the aid of a grant from the International Development Research Centre (IDRC), Canada and the Department for International Development (DFID), UK.

Abstract

As people become increasingly mobile and the urban traffic patterns become more complex there is an emerging need for the transport planning to become a truly continuous process. Developing countries that continue to rely mainly on infrequent and expensive survey data for transportation planning and management, will struggle to cope with this changing dynamic. Mobile network big data provides a promising avenue that can potentially fill this gap without requiring significant investment in sensor networks. This paper builds on earlier work to explore in greater depth the potential of mobile network big data (MNBD) to provide high value mobility insights that can support a continuous approach to transport planning. We evaluate the levels of accuracy and detail that mobility insights based MNBD can deliver by comparing multiple recently developed approaches for estimating mobility and validating the results against the data from traditional transport forecasting methods. We discuss inherent limitations of MNBD in generating insights for transport planning and propose various methods to address these limitations in future work. The value of such work is that this extends the state of the art of using such new data sources that can transform traditional transportation planning.

Keywords: Big Data, Call Detail Records, Transport Forecasting, Origin-Destination Matrices

1 Introduction

Some of the main insights used for transportation analyses and planning are those derived through a set of established analytical methods using data collected by means of surveys. This transportation forecasting framework is designed to leverage relatively small samples of rich mobility information to forecast transport demand. In contrast Mobile Network Big Data (MNBD) represents a much larger volume of data that accounts for a large fraction of the population. Mobility analysis using MNBD quite often provide output that correspond to some of the intermediate or final results of the traditional forecasting framework, based on significantly different intuition and analytical methods. These outputs provide greater temporal and spatial granularity than traditional forecasts. However additional information collected through surveys such as the trip motivations, vehicle mode choice etc. that enrich the output of the traditional forecasts is absent in MNBD based forecasts.

The traditional transport forecasting approach used in Sri Lanka is predominantly based on the well-known four-step model, which consists of trip generation, trip distribution, mode choice, and route assignment, in that specific order.

Trip generation involves identifying the volumes of trips in and out of each traffic analysis zone. These volumes are derived based on household-level travel surveys, which also incorporate numerous socio-economic elements and land usage forecasts. The resulting trips are identified by the associated purpose such as work or leisure. Once trips have been identified they are distributed among pairs of traffic analysis zones and adjusted for population in a process known as trip distribution. The resulting output of this stage is usually a set of Origin Destination (O-D) matrices, which represent the trip distribution. Each data point in an O-D matrix represents the volume of trips from a source zone (Origin) to a destination zone (Destination) during the relevant time period. Gravity models, which distribute trips based on the population of the zones, distance between the zones, etc., are a common technique used at this stage. In the third stage, the different modes of travel such

as private automobile, bus, train, etc. are predicted and assigned for the trips identified in the previous stages. Discrete choice modeling with the different transport modes as choices is a typical approach used at this stage. The fourth and final stage of the classical transport forecasting approach is route assignment. This involves assigning the identified trips to the road network based on origin and destination. This is commonly known as the route choice or route assignment. The task is complicated by the inter-dependency between the transport demand at a given time and the travel time between the origin and destination by a given route.¹

The frequency of these traditional forecasts as well as the degree of accuracy and detail depends entirely on the travel surveys that provide the source data. Unlike other surveys such as the Household Income and Expenditure Survey (HIES; carried out once every 4 years) and the decennial National Census, travel surveys are not carried out at regular intervals. In general a new travel survey is carried out on an as needed basis such as in the case of an extensive travel survey in 2013 for the purpose of developing a Western Province master plan. The driving cause for limiting travel surveys in this form is the cost and effort involved. The travel survey in 2013 funded by JICA is estimated to have cost approximately USD 400,000².

As people become increasingly mobile and the urban traffic patterns become more complex there is an emerging need for the transport planning process (of which the traditional transport forecast is a part) to become a truly continuous process. Continuous monitoring of urban mobility patterns and analysis of the impact of enacted policies requires a level of data extraction and analysis that is simply beyond the abilities of a traditional survey based mechanism. Mobile network big data provides a promising avenue that can potentially fill this gap without requiring significant investment in sensor networks.

Lokanathan, de Silva, Kreindler, Miyauchi, & Dhananjaya (2014) have explored the potential for MNBD to enable the process of continuously generating high value transport related insights in Sri Lanka. This paper handles the next logical step of more precisely exploring the level of detail, accuracy and the inherent limitations of the mobility insights that can be generated using MNBD through different approaches.

In this paper we discuss how the output from mobility analysis based on MNBD can be aligned with the different stages of the traditional forecasting framework familiar to transport planners and policy makers. We point out that leveraging the MNBD output within the traditional forecasting framework can lead to high quality insights that combine the strengths of both traditional forecasting and MNBD analysis.

2 Literature Review

Mobile network operators collect different types of spatio-temporal data associated with their customers in the daily course of operations. These include Visitor Location Register (VLR) data which generates a record whenever the base station serving a subscriber changes and Call Detail Records (CDR) where a record is generated when a subscriber uses a service such as taking/receiving a call. VLR records represent a high resolution dataset of human mobility that is independent of the use of mobile phone. However due to the extremely

¹ For a more thorough treatment of the four-step model refer to McNally (2008)

² Estimated based on interviews

large volumes of data involved and the associated storage and processing costs, using VLR for analyzing human mobility at large scale is difficult. In contrast CDRs which represent a more manageable form that retains many of the advantages of VLR albeit to a lesser extent has been used extensively for understanding transport in developed and developing countries.

Recent research using MNBD has introduced numerous methods of mobility analysis with different levels of complexity. Some of these recreate insights and outputs generated from a traditional forecasting approach at greater spatio-temporal resolutions. Others generate wholly new insights only possible with MNBD.

Several approaches for estimating O-D matrices based on CDR have been proposed with encouraging results using datasets from the United States, Bangladesh, Spain etc. (Bahoken & Raimond, 2012; Calabrese, Di Lorenzo, Liu, & Ratti, 2011; Jiang et al., 2013; Iqbal, Choudhury, Wang, & González, 2014; Wang, Hunter, Bayen, Schechtner, & González, 2012; Caceres, Wideberg, & Benitez, 2007). Wang et al. (2012) extends an O-D analysis for Boston and San Francisco bay area to that of vehicle flow and congestion analysis using additional information. There have also been attempts made to identify mode of travel of individuals based on their call detail records (Doyle, Hung, Kelly, McLoone, & Farrell, 2011; Wang, Calabrese, Di Lorenzo, & Ratti, 2010).

Lokanathan, de Silva, Kreindler, Miyauchi, & Dhananjaya (2014) carried out preliminary analyses of MNBD from Sri Lanka to understand human mobility. Their work articulated aggregate daily mobility patterns and spatio-temporal population density changes.

3 Data Source

The paper uses 13 months of Call Detail Records (CDRs) for nearly 10 million SIMs from multiple operators in Sri Lanka³. The data is completely pseudonymized by the operator i.e. the phone numbers have been replaced by a unique computer generated identifier. The researchers do not maintain any mapping information between the generated identifier and the original phone number.

Each CDR corresponds to a particular subscriber of the operator's network and is created every time a subscriber originates or receives a call. In the case of an in-network call (i.e. both parties on the call were subscribers in the same mobile network), two records are generated, one for each party. Each record contains the following attributes:

- Call direction: A code to denote if the record is an incoming or outgoing call
- Subscriber identifier: Anonymized identifier of subscriber in question
- Identifier of the other party: Anonymized identifier of the other party on the call
- Cell identifier: an ID of the cell (i.e. antenna) that the subscriber was connected to at the time of the call
- Date and time that the call was initiated
- Duration of the call

³ Due to the agreements with the operator we are unable to name the operators and cannot give a precise figure for the number of SIMs that were analyzed.

4 Research Methodology

Origin-Destination matrices are a key intermediate output of the traditional forecasting approach that serves as input for a variety of analyses including the latter stages of traditional forecasting framework. Traffic studies which estimate vehicle/passenger flows, congestion etc. in particular are benefited by accurate high resolution O-D matrices which quantify transport demand.

Recent research has introduced multiple approaches with different intuitions that can be employed to construct O-D matrices based on MNBD. The resulting O-D matrices need to be interpreted differently and given different needs have the potential to replace or supplement the traditional O-D matrices.

In this paper we consider three approaches for generating O-D matrices based on MNBD each with its own model of human mobility. We compare the results with the best available data derived using the traditional forecasting framework.

In analyzing mobility based on MNBD it is important to establish an appropriate unit of spatial resolution. Density of base stations in urban regions is generally higher than in the rural regions as mobile network operators construct towers to accommodate higher usage. Therefore the spatial resolution available in data for urban regions is higher than for rural regions. Additionally in urban areas where there's high mobile phone usage network operators either explicitly use network load sharing approaches or such load sharing is inherent in the network technology. Load sharing in a network results in different base stations serving the same geographical location at different times. This introduces uncertainty about the position of a user connecting to a base station. Spatial resolution selected for analysis should maximize the spatial resolution that can be retained considering the variation between urban and rural regions and minimize the uncertainty of user positions.

We employ a grid-based spatial coordinate system based on Williams, Thomas, Dunbar, Eagle, & Dobra (2014) with 1km^2 resolution to manage these considerations. The analyses considered in this paper are based on a single grid due to computational limitations even though Williams et al (2014) advocate using multiple grids with different positioning to derive averaged measures. The generated grid has been positioned to maximize the inter-tower distances between neighboring cells in the grid to minimize positioning errors.

The O-D matrix estimation techniques discussed below follow two general stages. The first involves breaking down the movement of individuals into distinct sequences of trips. In the second step the resulting trips are aggregated across defined origin and destination locations.

4.1 Stay Based O-D Matrix Estimation

The traditional survey approach to data collection for transport forecasting focus on the origin and destination of trips made by people, which makes it simpler to associate semantic information like motivations and understand travel patterns and motifs. Calabrese et al. (2011) and Jiang et al. (2013) suggest approaches for estimating O-D matrices that attempt to identify origins and destinations that have significant semantic value. Jiang et al. (2013) goes on to infer human activities associated with the identified trips utilizing land use survey data.

We employ an approach based on Calabrese et al. (2011) and Jiang et al. (2013) with slight modifications in an attempt to align the interpretation of the O-D matrix based on MNBD with that of the traditional forecast.

As the first step we identify instances a user has been stationary at a geographical location with the associated time period, which are referred to as **stays**⁴ in the rest of the paper. In terms of CDRs we define a **stay** as a contiguous series of records such that,

- (1) Any two records in the series are less than a distance D apart, where $D = 1km$
- (2) The entire series of records should span a period of more than 10 minutes
- (3) Two contiguous records are separated by a time interval T such that $T \leq 1hour$

We use the base station locations at this stage of the analysis and the maximum diameter (D) of a stay controls the spatial resolution at which stays are identified. Similar to Calabrese (2013) maximum diameter of 1 km was selected as suitable for defining a locality to account for the uncertainty of location due to network load sharing. We take the medoid BTS location of the set of BTS involved in a **stay** as its representative location which is further quantized based on the $1km^2$ grid coordinate system. This approach is preferable to directly applying the grid coordinate system, since a locality defined by the BTS connected during a stay doesn't necessarily conform to the grid positions artificially imposed. A stay is represented as a 4-tuple, $\langle user-identifier, location, start-time, end-time \rangle$.

Each consecutive pair of **stays** during a day for an individual is considered as the origin and destination of a trip. This means that if there are n **stays** identified for an individual during a day, $n - 1$ trips would be captured. A trip is represented as 5-tuple, $\langle identifier, origin, destination, final-origin-time, first-destination-time \rangle$.

O-D matrices are estimated by aggregating the trips identified for every person in the dataset at the grid coordinates that represent trip end points. We generated hourly and daily O-D matrices as well as for each day of week. In generating hourly O-D matrices trips can be partitioned based on either the last recorded time at the origin or the first recorded time at the destination. In contrast to Calabrese et al. (2011), we have chosen the first recorded time at the destination for this purpose as this reflects the earliest indication of a trip having occurred. However since the first recorded time at the destination is represents an upper bound for the period during which the trip actually occurred, O-D flows will be shifted forward in comparison to actual motion. We consider this distortion to be insignificant for our purposes.

4.2 Transient O-D matrix estimation

The mobile phone usage of individuals is not restricted to locations where they remain stationary for a period of time or have significant meaning to them. Mobile phones are quite frequently used during travel at intermediate or transient locations between the starting point and the intended destination. The stay based approach focuses on trip endpoints where individuals have remained stationary for a period of time and potentially has significant meaning in their daily routine. As a result the approach tends to ignore the

⁴ The term stay is based on "stay point" and "stay region" by Jiang et al. (2013). However our definition of a stay utilizing mobile data is more closely related to the definition provided by Calabrese et al. (2011)

mobility information present as transient locations. Wang et al. (2012) propose an approach that maximizes the mobility information extracted from CDR's by making use of transient locations. This approach has been applied in San Francisco Bay and Boston in the United States as well Dhaka in Bangladesh (Iqbal et al., 2014). We adopt the same technique with minor modifications to the CDR data from Sri Lanka.

In this approach a trip is identified from the CDRs by a consecutive pair of records such that,

- (1) The records indicate a displacement, i.e. the BTS-es utilized for each record is different
- (2) The records are separated by a time interval T_{Interval} where,
 $10 \text{ minutes} \leq T_{\text{Interval}} \leq 1 \text{ hour}$

In contrast to Wang et al. (2012) we introduce an additional temporal constraint of a minimum time interval between consecutive records of 10 minutes. This change accounts for the high degree of uncertainty in mobility perceived through CDR during short periods due to stationary individuals connecting to different neighboring towers by ignoring such records for identifying trips. The upper bound of 1 hour for the allowed time interval between consecutive records controls the temporal resolution of the generated trips as well as the degree of alignment with actual trips made by individuals.

A trip is represented as 5-tuple similar to the previous approach, $\langle \text{identifier, origin, destination, origin-time, destination-time} \rangle$

The trips are aggregated at the grid coordinate level to estimate O-D flows. The time recorded at the destination was used when partitioning the flows into time intervals within a day with the same intuition as in the stay based approach.

4.3 Frequent trip based approach

Human mobility demonstrates a high level of temporal and spatial regularity which has been validated with MNBD among other data sources (González, Hidalgo, & Barabási, 2008). Traditional transport forecasts utilize this intuition through collecting data on regular mobility behavior and associated travel motivation etc. Both stay based and transient approaches for estimating O-D matrices make no distinction between the regular and ad-hoc elements of mobility of individuals. We applied an approach that attempts to capture the regular element of mobility through MNBD based primarily on Bayir, Demirbas, & Eagle (2009).

Bayir et al. (2009) propose first identifying daily location sequences for individuals bounded by end-locations (which have a similar intuition to **stays**) that also include transient points. Location sequences that occur frequently in the daily routine of an individual are then identified based on a frequency threshold.

In our variation of the method we simplify the analysis by foregoing the identification of specific location sequences bounded by end locations. Instead we consider the entire daily sequence of locations after filtering out consecutive occurrences of the same location. We identify frequent location sequences from the daily sequences for an individual through the application of an efficient frequent sequence mining algorithm. A location sequence is deemed frequent if it occurs on at least 10% of the daily sequences of an individual. The process of identifying frequent locations only enforces the order of occurrence of locations and doesn't require a sequence to be contiguous. While frequent sequences of more than

two locations can provide insights into route choice of an individual we limit the analysis to sequences of two locations to understand O-D flows.

We use frequent sequences of length two for each individual as the endpoints of trips to be observed. We count the trips for individuals separately for each day of the week and each day is separated into four time periods that we expect corresponds to different mobility behaviors, proposed by Wang et al (2012).

Time Periods used for the analysis are,

1. Morning (6am - 10am)
2. Afternoon (10am - 4pm)
3. Evening (4pm - 8pm)
4. Night (8pm - 6am)

The two shorter periods Morning and Evening focus on daily commuting hours while Night and Afternoon periods align with regular Home and Work hours.

As outlined in Equation 1, the trip count calculated for each trip $Trip_i$ for an individual i for a period $Period_k$ where $k = \{Morning, Afternoon, Evening, Night\}$ for a day D_j where j is a value from 0 to 6 representing the day of the week, is then converted to an estimate of the probability P of the trip being made by the individual during that day-period combination.

$$P(Trip_i|D_j,Period_k) \approx \frac{\text{Frequency of } Trip_i \text{ during } D_j \text{ and } Period_k}{\# \text{ of times } i \text{ had at least 1 record during } D_j \text{ and } Period_k}$$

Equation 1: Probability of performing a trip on a particular day-period combination

This estimation assumes that communication behavior as evidenced by the frequency of call and internet usage is independent of mobility.

4.4 Comparison with output from traditional transport forecasting

The best available validation data was trip generation counts at the Divisional Secretariat Division level (DSD) for the Western Province except Colombo and Thimbirigasyaya DSDs (data for these two DSDs was not available in a usable format). We compared the trip generations estimated based on the three MNBD methods by aggregating the results at DSD level. The O-D results for the transient approaches were based on 13 months of CDR records while the results for the stay based approach was based on 1 month of data due to time and storage constraints. A weighted least square linear regression was run in each case against the traditional forecast where the weight equaled $\frac{1}{MNBD \text{ estimate}}$.

As revealed by the three figures (Figure 3, 4, & 5 in Annex 1), the three approaches produce estimates that have a high-level of correlation with the traditional forecasting figures. The frequent trip approach seems to be marginally better than the other two approaches.

5 Results and discussion

The stay-based approach captures only a very small number of trips compared to the transient and frequent trip approaches due to the strict spatio-temporal constraints. They also bias results towards those of subscribers with high mobile activity, which may not be

reflective of the broader population. For example mobile activity is lower in rural or remote areas than in dense urban environments. However for urban regions with generally higher mobile usage activity by residents, the approach is very useful for identifying stays.

The ability of the transient approach to generate hourly or daily O-D flows makes it particularly suitable for short-term mobility analysis. This type of output is not possible with the traditional forecasting framework due to the time and effort required to collect and process the data. Figure 1 shows the circadian rhythm of mobility for Sri Lanka extracted from the transient approach.

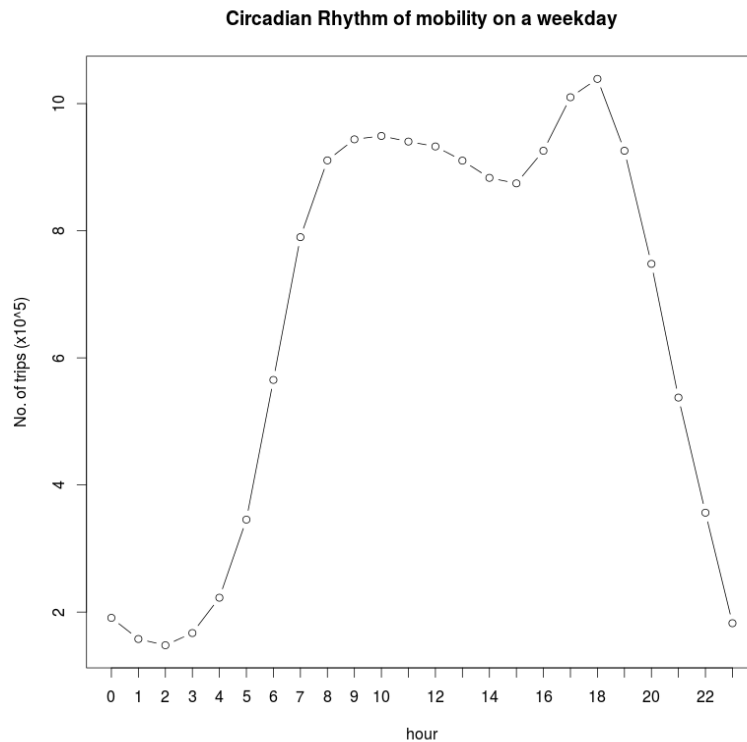


Figure 1: Diurnal Pattern of mobility from transient approach

The stay based approach and the transient trip approach impose strong temporal constraints in identifying trips. This has two shortcomings. In the first place, long distance trips may not be identified at all. Secondly the likely result is that actual long distance trips get falsely detected as a series of shorter trips, but without any ability to regenerate the actual long distance trip that had occurred. The frequent trip approach doesn't impose such constraints and is therefore likely to capture all origins and destinations corresponding to a regular trip irrespective of the distance. It is worth noting that this approach ignores ad-hoc movement by individuals which when considered for the entire population may be significant. However as ad-hoc movement can't be interpreted in terms of travel motivations, this approach is useful where understanding travel motivations and motifs is the focus.

Given an area of interest transport planners use O-D estimates to understand the geographical distribution of trip origins and trip destinations. This allows them to identify major flows in and out of region and selectively manage transport demand through different treatments including building capacity on the road network or developing regions which may serve as alternative destinations for specific needs. Figure 6 shows the geographical

distribution of origins of trips in to a region within Fort, the commercial hub of Colombo and Borella junction within the largest suburb of the city using O-D flows generated with the frequent trip approach. Significant sources of trips to the Fort region are distributed along the coastline aligning with the A2 highway to the south and A3 highway to the north. In contrast the significant sources of trips to Borella junction are bounded within a smaller region that extends more into the interior of the Colombo district.

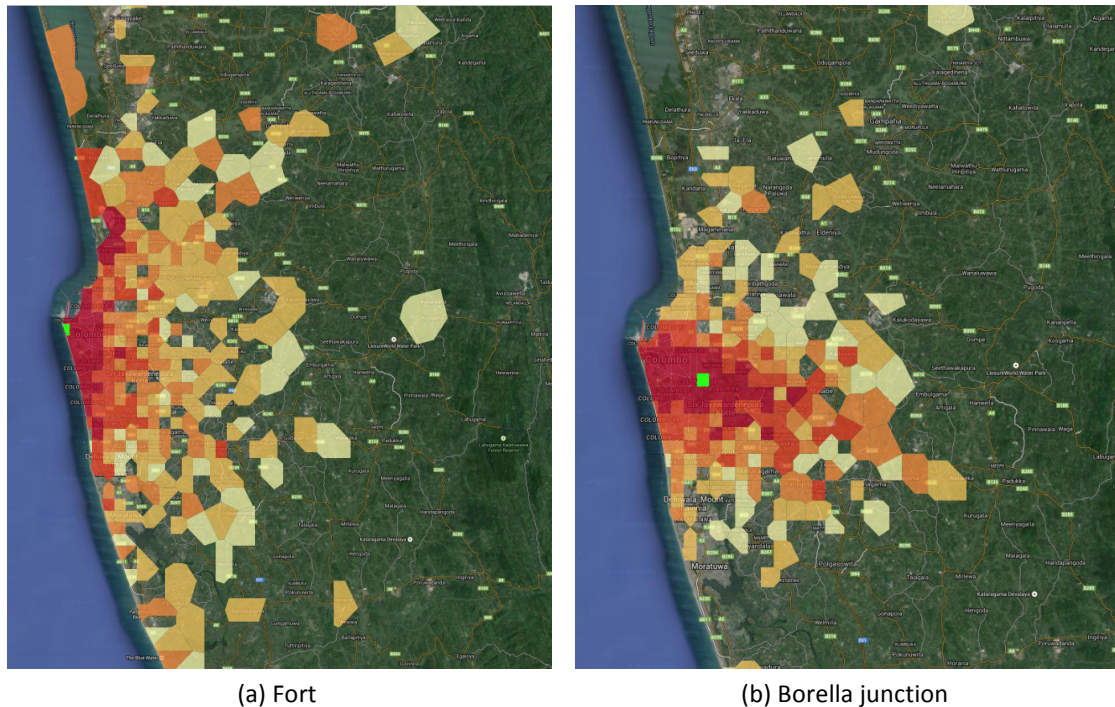


Figure 2: Geographical distribution of trip origins

5.1 Utilizing MNBD outputs within a traditional forecasting framework

While MNBD based mobility analysis provides outputs with high levels of spatio-temporal resolutions, these outputs have weaknesses that limit their use without further processing. These include the lack of any socioeconomic, demographic, travel motivation data as well as numerous sampling biases.

O-D flows estimated with MNBD represent only a fraction of the actual flows since only the movements of mobile users are captured. These estimates can be calibrated with the latest available traditional forecasts to derive realistic figures from MNBD. MNBD based O-D estimates can be derived continuously which will capture any changes in transport patterns until the next forecast based on the traditional approach. This allows the transport forecasts during the period between traditional forecasts remain more relevant and up to date.

MNBD based mobility analysis is vulnerable to a number of sampling biases. By the very mode of generation of CDR, the resulting data only captures the movement of mobile users. Even among mobile users mobility analysis techniques are more likely to capture the mobility of active users more clearly compared to less active users. The insights on mobility derived from MNBD are further affected by the overall mobile penetration, as well as the market shares of the operators whose data is analyzed. In order to correct for these biases we need to understand the representativeness of the data in terms of the different estimated variables such as resident mobile population against corresponding official figures derived from census etc. For example we can adjust O-D flow estimations for mobile phone

penetration and operator market share by scaling flows originating at a given location by the ratio between the census population and the resident mobile users. Adjustment of mobility estimates for bias towards active mobile users requires a more complex approach that models how the mobility estimation for an individual varies with activity variables such as daily call count and average inter-call period etc.

The surveys done as a part of the traditional forecasting approach collect a wealth of demographic, socioeconomic and travel motivation information that can't be directly extracted from the MNBD. Arai, & Shibasaki (2013) discuss a supervised machine learning technique that associates the demographic parameters from the surveys with the mobile users in the CDR dataset by using the mobility variables present in both.

6 Challenges and future work

6.1 Sparsity of data

Sparsity of CDR records for an individual within a day is one of the primary issues faced when attempting to ensure representative estimate of mobility. In our dataset 90% of the users have less than 25 daily records. In addition these records are not evenly distributed during the day and this further reduces the amount of mobility information that can be extracted. However we can use the fact that the individuals don't necessarily always use their mobile phones from the exact same locations. We can use records of similar day and time periods together to derive a better understanding of mobility by assuming a high degree of regularity of individual mobility based on work by González et al. (2008) and controlling for infrequent ad-hoc mobility. It is proposed that we can develop enhanced CDR dataset where locations for times no records are present for a user are predicted using a probabilistic model based on this rationale.

6.2 Travel motifs

In developing O-D flow estimates we have not made any attempt to understand the activity associated with individual trips. The traditional forecasting approach employs motivations for travel and the activities performed at locations to develop a semantic model of travel motifs. Transport planners use this understanding combined with the geographical distribution of facilities to calibrate the O-D flows as well analyze mobility between regions by cause.

Lokanathan et al. (2014) has carried out a preliminary analysis of human mobility in Sri Lanka by assuming a single "Home-Work-Home" travel motif. We plan to carry out a more detailed analysis to understand the travel motifs of individuals and the key travel motifs driving aggregate mobility.

7 Conclusion

MNBD based mobility analysis represents an opportunity to provide enhanced transport forecasts with greater temporal and spatial resolution that retain their validity throughout. In particular developing nations like Sri Lanka, which generate forecasts, based on expensive and infrequent surveys, a forecasting framework that combines MNBD analysis with traditional forecasting can deliver high quality outputs that benefit from the strengths of both. This paper critically evaluates that potential, comparing three different approaches for

extracting human mobility patterns using MNBD and how that can be integrated and used in conjunction to the traditional forecasting approach.

Developing a framework that effectively leverages both MNBD and traditional forecasting data, requires further work. This would firstly require a few enhancements to overcome the limitations of data sparsity inherent in CDR data. Secondly there is further investigation required on integrating travel motif into the analyses of MNBD so that it can more closely align with current semantic models used by transportation planners.

As people become increasingly mobile and the urban traffic patterns become more complex there is an emerging need for the transport planning to become a truly continuous process. Hence extending the state of the art as this paper does, will be critical, if transportation planning is to adapt to this changing dynamic.

8 References

- Arai, A., & Shibasaki, R. (n.d.). Estimation of Human Mobility Patterns and Attributes Analyzing Anonymized Mobile Phone CDR: Developing Real-time Census from Crowds of Greater Dhaka. *Ceur-Ws.Org*. Retrieved from <http://ceur-ws.org/Vol-1136/paper2.pdf>
- Bahoken, F., & Raimond, A. M. O. (2013). Designing Origin-Destination Flow Matrices from Individual Mobile Phone Paths: The effect of spatiotemporal filtering on flow measurement. In *ICC 2013-International Cartographic Conference*.
- Bayir, M. A., Demirbas, M., & Eagle, N. (2010). Mobility profiler: A framework for discovering mobility profiles of cell phone users. *Pervasive and Mobile Computing*, 6(4), 435–454. doi:10.1016/j.pmcj.2010.01.003
- Caceres, N., Wideberg, J. P., & Benitez, F. G. (2007). Deriving origin–destination data from mobile phone network. *IET Intelligent Transport Systems*, 1(1), 15. doi:10.1049/iet-its:20060020
- Calabrese, F., Di Lorenzo, G, Liu, L., Ratti, C. (2011). Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area.
- Doyle, J., Hung, P., Kelly, D., McLoone, S., & Farrell, R. (2011). Utilising mobile phone billing records for travel mode discovery. In *ISSC 2011*. Dublin: Trinity College Dublin. Retrieved from <http://eprints.nuim.ie/3649>
- González, M. C., Hidalgo, C. a, & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196), 779–82. doi:10.1038/nature06958
- Iqbal, M. S., Choudhury, C. F., Wang, P., & González, M. C. (2014). Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40, 63–74. doi:10.1016/j.trc.2014.01.002
- Jiang, S., Fiore, G. A., Yang, Y., Ferreira, J., Frazzoli, E., & González, M. C. (2013). A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and

Opportunities. In *Proceedings of 2nd ACM SIGKDD International Workshop on Urban Computing*. Chicago, IL.

Lokanathan, S., Silva, N. de, Kreindler, G., Miyauchi, Y., & Dhananjaya, D. (2014). *Using Mobile Network Big Data for Informing Transportation and Urban Planning in Colombo*. (2014). Available at <http://dx.doi.org/10.2139/ssrn.2526642>

McNally, M. G. (2008). The four step model. In Hensher & Button (Eds.). *Handbook of Transport Modelling*. Pergamon, 2nd Edition.

Wang, H., Calabrese, F., Di Lorenzo, G., & Ratti, C. (2010). Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *13th International IEEE Conference on Intelligent Transportation Systems* (pp. 318–323). IEEE. doi:10.1109/ITSC.2010.5625188

Wang, P., Hunter, T., Bayen, A. M., Schechtner, K., & González, M. C. (2012). Understanding road usage patterns in urban areas. *Scientific Reports*, 2, 1001. doi:10.1038/srep01001

Williams, N. E., Thomas, T. A., Dunbar, M., Eagle, N., & Dobra, A. (2014). Measures of Human Mobility Using Mobile Phone Records Enhanced with GIS Data, 33. Retrieved from <http://arxiv.org/abs/1408.5420>

Annex 1: Validation of MNBD estimates with traditional forecasting data

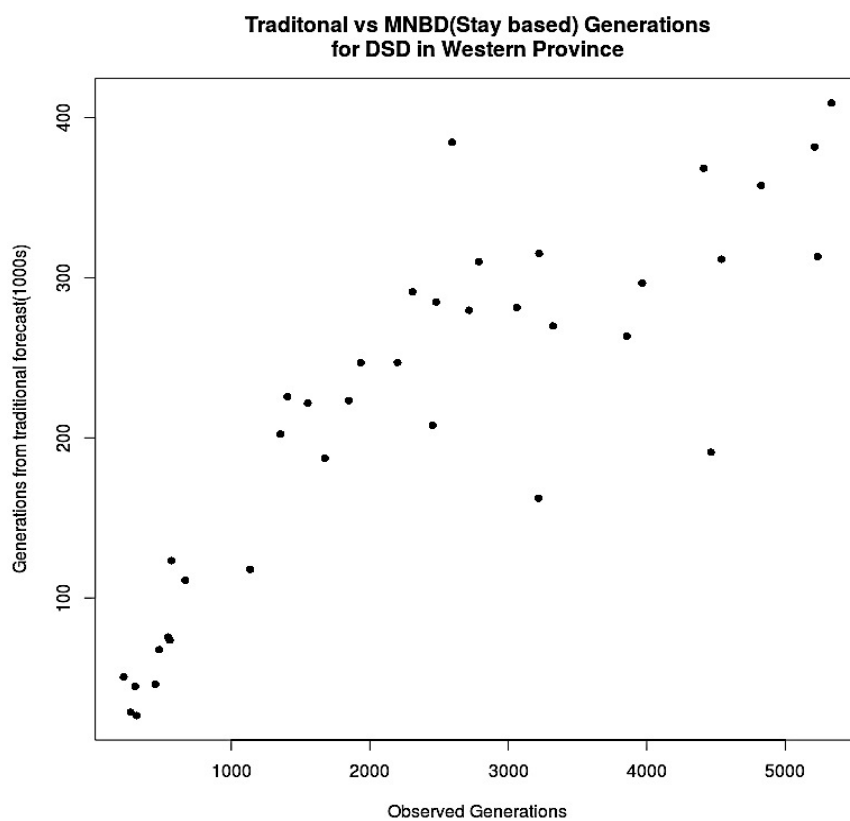


Figure 3: Relationship between trip generations estimated with stay based approach and the generation figures from traditional forecast

Weighted residuals:

Min	1Q	Median	3Q	Max
-280.29	-91.78	-20.97	105.44	391.62

Residual standard error: 1334 on 35 DF
Multiple R-squared: 0.824
Adjusted R-squared: 0.819
F-statistic: 163.9 on 1 and 35 DF
p-value: < 9.1777e-15

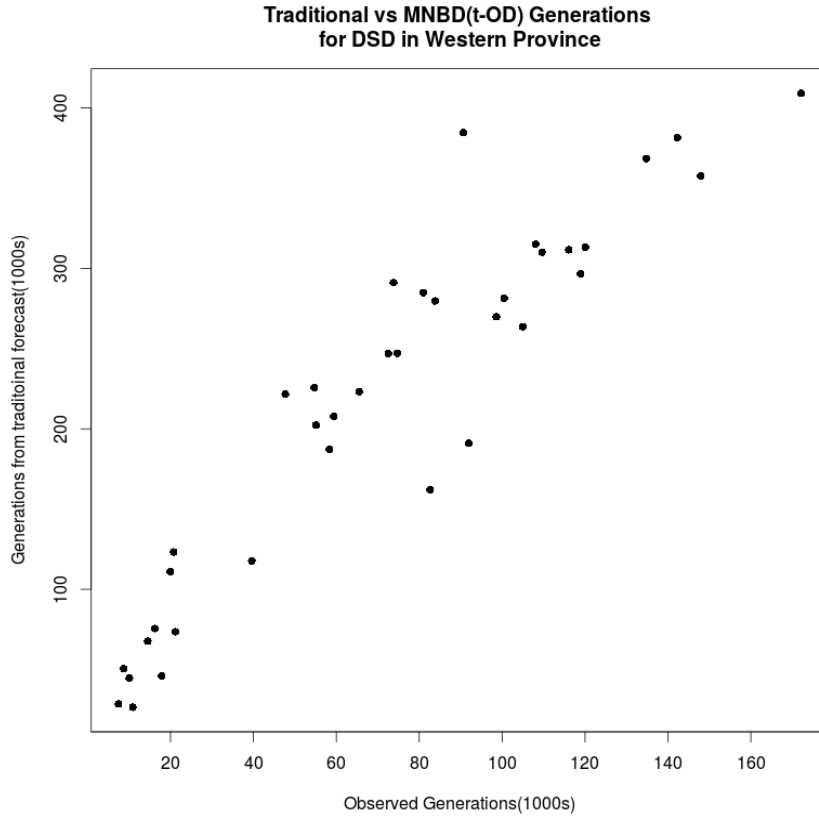


Figure 4: Relationship between trip generations estimated with transient approach and the generation figures from traditional forecast

Weighted residuals:

Min	1Q	Median	3Q	Max
-280.29	-91.78	-20.97	105.44	391.62

Residual standard error:	172.2 on 35 DF
Multiple R-squared:	0.9059
Adjusted R-squared:	0.9032
F-statistic:	337 on 1 and 35 DF
p-value:	< 2.2e-16

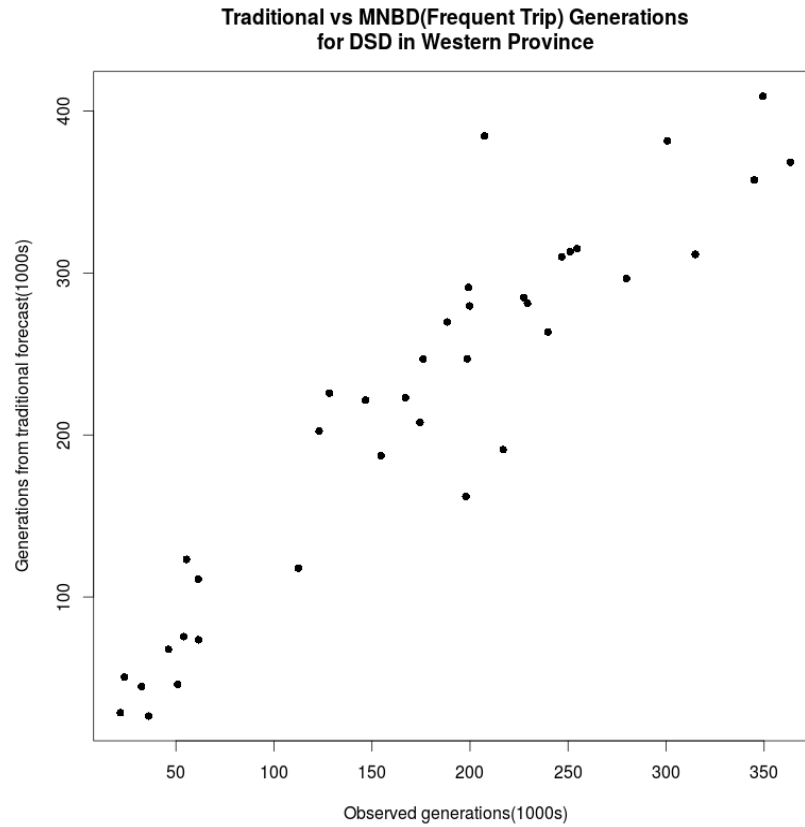


Figure 5: Relationship between trip generations estimated with frequent trip approach and the generation figures from traditional forecast

Weighted residuals:

Min	1Q	Median	3Q	Max
-183.83	-77.097	1.995	67.581	284.418

Residual standard error: 104.3 on 35 DF

Multiple R-squared: 0.9118

Adjusted R-squared: 0.9093

F-statistic: 361.8 on 1 and 35 DF

p-value: < 2.2e-16