ORIGIN-DESTINATION MATRIX ESTIMATION FOR SRI LANKA USING MOBILE NETWORK BIG DATA

Danaja Maldeniya and Sriganesh Lokanathan, LIRNEasia, Sri Lanka

Amal Kumarage, University of Moratuwa, Sri Lanka

Abstract: In this paper we discuss the potential of mobile network big data (MNBD) for providing insights for transportation planning that can supplement traditional methodologies. Adapting recently developed techniques for analyzing Big Data, we demonstrate their applicability to the Sri Lankan context. The paper assess the efficacy of two different methodologies for providing insights related to transportation planning from MNBD, and we validate the results using data from traditional transportation forecasting methods. We discuss the resultant limitations and the issues of attempting to leverage such data and propose various methods to address the constraints.

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

1. INTRODUCTION

Understanding patterns of human mobility is a requisite precursor not just for developing an effective transportation policy, but also for strategic planning and management of the extant transportation system. Traditional transport forecasting techniques depend on surveys. Cost and scale implications for these surveys mean that compromises have to be made in term of the level of spatial and temporal granularity of the data. Developed economies on the other hand, fueled by the increased 'datafication' from the greater use of sensors can avail of a host of additional as well as high-frequency data to feed into the transportation planning systems. These benefits have not uniformly flowed to developing economies, which still depend heavily on traditional survey instruments to fulfill many of their data requirements.

On the other hand developing nations (like developed economies) do exhibit signs of ubiquitous telecommunication access, primarily through mobile phones. Sri Lanka for example has a mobile penetration of over 90% and coverage of nearly 100% of the landmass (Telecom Regulatory Commission of Sri Lanka, 2014). Due to the nature of mobile phone technology, which requires mobile phones to connect with strategically placed base-stations, the mobility of mobile phones can be considered a proxy for human travel patterns.

As a result Mobile Network Big Data (MNBD) affords the possibility of obtaining spatially fine and high frequency data on human mobility, especially in urban areas, which have a higher density of base stations (and therefore greater spatial granularity in the data).

Recent research (Calabrese, Di Lorenzo, Liu, & Ratti, 2011; Wang, Hunter, Bayen, Schechtner, & González, 2012; Jiang, Fiore, Yang, Ferreira, Frazzoli, & González, 2013) utilizing MNBD for understanding and forecasting urban mobility has introduced methods of varying complexity and dealing with different aspects of mobility. Often each application of the method has been to a single country (or region) using only one operator's data. Whilst these methods have shown applicability of MNBD for providing transportation related insights in those specific countries, what is less clear is the generalizability of the proposed methods to other regions and/ or the performance/ effectiveness of each of the different methods to the same underlying data.

The aim of this paper is two-fold. Firstly it attempts to adapt the methods developed using datasets from other countries for the case of Sri Lanka (specifically the Western Province). Secondly it aims to test the efficacy of different methods to derive data (currently obtained only through infrequent surveys) relevant for transportation planning, management and policy. To our knowledge there is no prior work that has done so, and also compared the resultant insights with those from traditional survey instruments.

2. THE FOUR STEP MODEL

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 result 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]

3. STATE OF THE ART

Depending on their sophistication, mobile networks capture a range of spatio-temporal data that can be valuable for transportation planning. Of these, the most common, and one that is universally captured by all mobile operators is passive positioning data in the form of Call Detail Records (CDRs). CDRs are generated automatically by the network and captured in the operator’s logs for billing, and network management purposes (from keeping track of the handset in relation to its network elements, to understanding network load).

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.

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

Call Direction	Calling Party Number	Called Party Number	Cell ID	Call Time	Call Duration
1	A24BC1571X	B321SG141X	3134	13-04-2013 17:42:14	00:03:35

Table 1: Structure of a pseudonymized Call Detail Record (CDR)

Table 1 is a stylized representation of the elements of a pseudonymized CDR.[2] *Call Direction* defines whether the record is for an incoming or an outgoing call. *Calling Party Number* and *Called Party Number* are each pseudonymized references to a subscriber. The *Cell ID* field provides the reference to the cell (i.e. antenna) to which the relevant mobile phone connected for the recorded activity. Depending on the directionality of the *Call Direction* variable, the Cell ID field corresponds to either the caller or callee. Each cell is associated with a corresponding Base Transceiver Station (BTS), which in turn has a latitude and longitude associated with it. A BTS may (and often does) host more than one antenna. The *Call Time* field provides the time at which the recorded activity was initiated.

The spatio-temporal data derived from CDRs can be used to directly estimate trip distribution between zones. This is in contrast to the four-step model where trip distribution is the second step. However data derived in this manner from MNBD can require numerous adjustments, including correcting for selection bias, and scaling to the actual population.

The key element in analyzing trip generation and distribution based on MNDB is the identification of movement trajectories or trips of individuals. Utilizing CDR data from the Boston metropolitan region, Calabrese et al. (2011) suggest an approach for identifying individual trips using a number of temporal and spatial constraints to minimize the impact of localization errors. Their approach collapses event locations from the CDRs for an individual into a contiguous series of virtual locations. Each virtual location represents a contiguous series of events, where the event locations are not more than 1km apart, and the time between the events is less than 10 minutes. The centroid of all antenna locations associated with that particular virtual location, represents the position of the virtual location. The daily trajectory of an individual is thus represented by a chronological order of *virtual locations* and each non-identical consecutive pair of virtual locations corresponds to a trip. The trips are aggregated based the virtual origins and destinations at different temporal windows to derive O-D matrices. Jiang et al. (2013) discuss a similar technique, which identifies stay locations based on a roaming distance of 300m which is the maximum distance between any two antenna locations and a minimum separation of 10 minutes between the first and last records. This method uses a grid clustering approach to further aggregate the stay locations to stay regions for more coarse-grained analysis of travel motifs, and preferential return and exploration characteristics of individual mobility.

With both methods, there is an inherent uncertainty whether the derived location represents the two ends of a trip (i.e. origin or destination) or is actually an intermediate location during a trip. The approaches suggested by Calabrese et al. (2011) and Jiang et al. (2013) focus on aligning the origins and destinations of the trips derived from the analysis to the actual trip endpoints. Wang et al. (2012) discuss a different approach that focuses on maximizing the captured mobility information, and placing less importance on whether the captured locations represents end-points or intermediate points, of a particular trip. In this approach any two consecutive non-identical locations in the daily location sequence of an individual, which occur within 1 hour of each other are considered the origin. Due to more relaxed spatial and temporal constraints this approach captures more mobility information from the CDRs as compared to the two earlier

[2] The actual CDRs contain the phone numbers of the subscribers. When obtained for research purposes such as this, the phone numbers are often replaced with a random identifier (or pseudonym) to maintain anonymity. The authors do not maintain any mappings the identifiers and the phone numbers they represent.

approaches. However this also means that the derived results are more sensitive to noise in terms of localization errors that may lead to over-estimation of very short distance trips.

Bayir, Demirbas & Eagle (2009) propose an alternate approach, which initially generates multiple mobility paths per person for each day. The paths are generated by incorporating both transient locations and trip end points, each of which are derived by methods similar to those articulated by Jiang et al. (2013) and Calabrese et al. (2011) respectively. Frequent mobility paths or trips are then identified for individuals through the application of a frequent-sequence mining algorithm, with suitable support and confidence parameters. The systematic O-D matrices generated in this manner can be considered estimations of the regular mobility or commuting patterns of the considered population.

Lokanathan, de Silva, Kreindler, Miyauchi, & Dhananjaya (2014) carried out a preliminary analyses of MNBD from Sri Lanka to understand human mobility. Their work articulated aggregate daily mobility patterns and spatio-temporal population density changes. However they did not explore fine-grained mobility patterns of the kind being conducted here.

4. THE DATASET

The study utilizes a month of historical and pseudonymized CDR data for nearly 10 million Subscriber Identity Modules (SIMs) obtained from multiple operators in Sri Lanka.[3] Collectively this dataset contains over 1.4 billion CDRs, with a volume of almost 120 Gb.

5. METHODOLOGY

Both 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. The O-D matrices estimated in this paper employ Voronoi cells defined for the mobile network base stations as the origin and destination locations.

5.1 Stay Based O-D Matrix Estimation

With some modifications, the stay-based approach utilizes techniques developed by Calabrese et al. (2011) and Jiang et al. (2013) that as a starting point, identifies stays. A stay for an individual is defined as a consisting of geographical location associated with a specific time period during which the individual was stationary.[4] In terms of the CDRs for an individual, a **stay** is identified by a continuous series of records such that,

(1) Two contiguous records in the series are less than a distance D apart, where D = 1km.
(2) Two contiguous records are separated by a time interval T_{Interval} such that
$10 \text{ minutes} \leq T_{\text{Interval}} \leq 1 \text{ hour}$

The maximum diameter (D) of a stay controls the spatial resolution at which stays are identified. Higher values for D result in lower spatial resolution with a larger upper bound for the area applicable for a stay. A 1km maximum diameter was chosen as an appropriate trade-off between the level of spatial resolution and the reduction of noise due to localization errors particularly in areas with very high tower density, similar to Calabrese et al (2011).

The CDR dataset demonstrated a relatively low level of individual activity level with 95% of the individuals having on average no more than 25 records per day. Given the relative sparsity of records and the intermittent nature of mobile phone activity for individuals, consecutive records

[3] The agreements with the operators do not allow us to mention the name of the operators, nor the precise number of SIMs whose data was analyzed.

[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)

for an individual may be separated by a significant time interval, which may hide actual motion. Neither Calabrese et al. (2011), nor Jiang et al. (2013) consider an upper bound for time interval We have introduced an additional temporal constraint of the time interval being less than 1 hour to mitigate the effect of hidden motion on the duration of the stays identified.

A stay is represented as a 4-tuple, <*user-identifier, location, start-time, end-time*>. The location field corresponds to the location of the medoid BTS of the set of BTS-es contributing to the **stay**.

The daily sequence of stays identified for an individual is used to generate the corresponding set of trips. Each pair of consecutive stays are considered the origin and destination of a trip and the last recorded time of the origin stay and the first recorded time of the destination stay are extracted as the relevant times. Therefore if **n** stays have been identified for an individual on a given day, this would represent a total of **n-1** trips for that day.

A trip is a represented as 5-tuple, <*identifier, origin, destination, final-origin-time, first-destination-time*>

Origin-Destination matrices are constructed by aggregating the identified trips at the base station locations. In estimating O-D flows for different periods of a day, trips can be partitioned based on either the final recorded time at the origin or the first recorded time at the destination. We chose the latter since it provides the first evidence of the trip having occurred. Since the first time recorded at the destination is an upper bound for the time at which the individual arrived at the destination, this approach can be expected to shift trip distribution forward in comparison to the actual movement.

5.2 Transient O-D Matrix Estimation

The transient based approach proposed by Wang et al (2012), has been applied to datasets from different countries (Bahoken & Raimond, 2013; Iqbal, Choudhury, Wang & González, 2014). We apply the same technique here to data from Sri Lanka, with small modifications.

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,
$$10 \text{ minutes} \leq T_{\text{Interval}} \leq 1 \text{ hour}$$

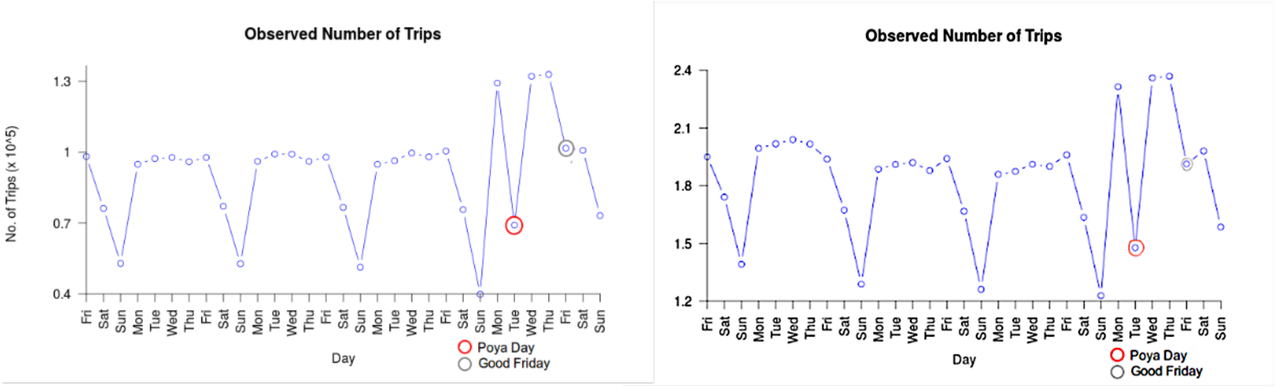
Unlike Wang et al. (2012), we also impose a temporal constraint of a minimum time (10 minutes) to the time interval between records. This change reduces the number of false displacements due to stationary individuals connecting to multiple neighboring towers during short time intervals, which may otherwise get captured as trips. The upper-bound limit of 1 hour on the time interval is used to manage the number of trips captured during a day for an individual, as well as the degree to which the trips identified through records may correspond to actual trips by the individual.

A trip is a represented as 5-tuple similar to the previous approach, <*identifier, origin, destination, origin-time, destination-time*>

The trips are aggregated at the BTS level to estimate O-D flows. The time recorded at the destination was used when partitioning the flows into time intervals within a day.

6. RESULTS

The two approaches outlined earlier were used estimate two sets of O-D matrices for the Western Province of Sri Lanka. Initially the O-D matrices were estimated at the lowest possible location resolution, which was at the BTS-level. The resulting O-D matrices were found to be sensitive to the false displacement effect where stationary individuals appear to be in motion due to connecting to different base stations in the vicinity. We generated a set of new origin and destination locations by generating 1km x 1km grid positioned to maximize distances between base stations in neighboring cells. The final O-D matrices were derived by aggregating BTS-level O-D matrices to the regions specified by the afore-mentioned grid.



(a) Stay Based Approach

(b) Transient O-D Approach

Figure 1. Daily total number of trips observed over a month

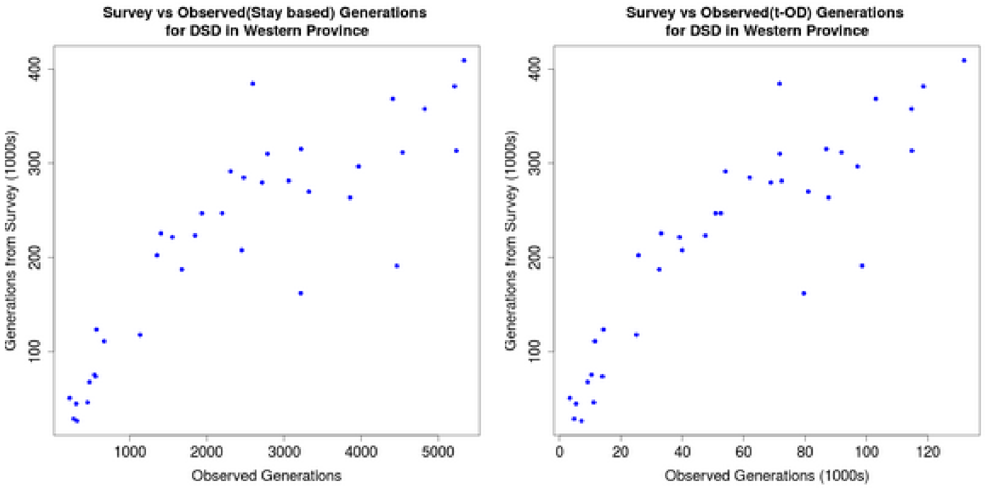
6.1. Stay Based O-D Matrix Estimation

Due to the strong spatio-temporal constraints applied in this approach we observe that stays are detected for only about 50% of the individuals having a CDR on a given day. Similarly trips are detected for only about 20% of the individuals having a CDR on a given day. We can conclude that stays and trips are more likely to be detected for individuals with high levels of activity (i.e. high number of CDRs) per day. What is less clear is high activity users are representative of the population and further research will be required.

6.2. Transient O-D Matrix Estimation

Unlike the stay-based approach, the transient approach utilizes more relaxed spatio-temporal constraints. This results in much better trip detection, with 45% of all individuals with a atleast one CDR on a given day, generating trips. Further more the average daily number of trips observed via this method is approximately 20 times greater than those generated using the stay based approach. However, the transient O-D matrices are more sensitive to false displacement in comparison to the stay-based approach due to the lack of a spatial constraint defining a locality. While the stay-based approach tends to capture travel between significant ‘end’ locations, the transient O-D matrices also capture the intermediate locations during travel, which possibly makes it more appropriate for traffic analysis.

6.3. Validation



Residuals

Min	1Q	Median	3Q	Max
-2777.2	-723.1	-128.6	1080	2961.9

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

(a) Traditional vs Stay based approach

Residuals

Min	1Q	Median	3Q	Max
-520.39	-144.13	-5.06	189.86	463.14

Residual standard error:	257.9 on 35 DF
Multiple R-squared:	0.8504
Adjusted R-squared:	0.8461
F-statistic:	198 on 1 and 35 DF
p-value:	5.272e-16

(b) Traditional vs transient O-D approach

Figure 3. Relationship between trip generations based on traditional and MNBD approaches

We aggregated the derived O-D matrices to the Divisional Secretariat Division (DSD) level, which is a third-level administrative unit in Sri Lanka. We compared our results for the Western Province (consisting of 40 DSDs) from each approach with DSD-level trip estimates generated from transportation surveys. A weighted least square linear regression was run in each case where the weight equaled $\frac{1}{MNBD\ estimate}$. The results of these comparisons for each approach are shown in Figure 3. The higher r-squared value for the results from the transient trip approach, suggest that this approach is marginally better than the stay based approach. However it is also noisier as compared to the stay based approach.

7. DISCUSSION

The results suggest that MNBD do capture underlying patterns of human mobility quite well. The transient trip approach is marginally better than the stay based approach when comparing the results to data from transportation surveys. Furthermore, the advantage of the former approach is that it is less biased towards users with high mobile activity. But this also means that it is more susceptible to noise in the form of localization errors. However the stay based approach does align more closely with actual origins and destinations of travel, which suggests that it is better for identifying congregations rather than for modeling mobility.

Irrespective of the approach, the results do highlight several issues and limitations that will need to be accounted for.

7.1. Accounting for Sampling Bias

Mobile network datasets are free of the human biases and errors inherent in observational data such as those obtained from surveys. However when utilizing CDR data, it is biased towards more frequent users of mobile phones. As the number of interactions an individual has with the mobile phone over the course of a day increases, it becomes possible to more closely map that individual's mobility, and the reverse is true as the number of interactions decrease. Therefore the issue of whether the aggregated mobility represented by the O-D matrices estimated with the CDR dataset, is truly representative of the entire observed population or just the more frequent users needs to be addressed for the results to become useful. Depending on the degree of bias identified it may be possible to correct O-D flows to be more representative or remove a fraction of the CDR data corresponding to activity levels that don't correlate with the majority of the observed population (Wang, Calabrese, Di Lorenzo, & Ratti, 2010)

7.2. Non-Uniform Tower Density

The BTS density strongly correlates with population density in a region. Therefore urban areas have a disproportionately larger faction of BTS-es than rural areas, when compared by landmass. For example Colombo District (a part of Western province) is the most urbanized region in Sri Lanka and has a significantly larger concentration of BTS-es than any other district in the country. As a result the spatial resolution that can be extracted from CDR data varies from region to region, with higher resolutions available in denser urban regions, and much lower resolutions in less dense rural regions. This means that an individual from a rural region with mobility similar to another in urban region will seem to be less mobile. Grid based approaches with fixed spatial resolution have been suggested as potential way to mitigate this issue (Williams, Thomas, Dunbar, Eagle & Dobra, 2014). The area of interest is split in to rectangles with identical dimensions based on the level of analysis being performed, for example 5 km square cells for district level flow of people or a more high resolution 2 km square cells for road network traffic analysis. Base stations that are situated within a cell are assigned to the center of the cell and cells that don't contain a base station are ignored in the analysis. One of the limitations of such an approach is that the quantization of space results in artifacts for events that are close to the borders of the cells. An individual calling from within a particular cell may in fact be connecting to a base station in a neighboring cell. Such artifacts can be accounted for by performing the analysis based on all possible grids at the same resolution by shifting the grid on the area of interest based on a suitable shifting distance and considering the average value of the measures being used. This approach can also mitigate the impact of a single location being served by multiple base stations at different times by assigning an average location to neighboring towers based on the considered spatial resolution.

7.3. Mapping Socio-Economic Data to Mobility

The O-D matrix estimation techniques discussed in this paper use the BTS as the origin and destination locations. Traditionally O-D matrices use traffic analysis zones that are created based on census data allowing a number of census based socioeconomic parameters to be associated with the O-D matrices. This ease of association simplifies the next steps in analysis required for transportation planning, including understanding travel motives, travel mode, and route choice. In the case of CDR based O-D matrices, the association of the census based parameters requires that the origin and destination locations used be mapped to traditional traffic analysis zones or that the census data be mapped to the more granular origins and destinations derived from MNBD.

7.4. Travel Mode

In estimating O-D matrices we have ignored the mode of travel as well the complications associated with vehicular traffic. A single vehicle may carry multiple individuals, each of whom may use a mobile phone during travel. The number of passengers, travel speed, choice of route and the physical dimensions vary depending on the type of vehicle. Therefore adjustments are needed to incorporate these elements so as to convert the O-D matrices that represent just human mobility to once that represent vehicular traffic (Wang et al., 2012).

7.5. Access to Data

Replicability of such research, and improvements (even incremental) will continue to remain a challenge so long as access to MNBD remains limited. However MNBD are inherently private data sets and it is not possible to make them fall under the purview of Open Data policies. Further more there can be competitive implications for the operators sharing such data. For example base station maps have business sensitivity. If it is known which operator’s data has been utilized, the analyses can reveal localized penetration numbers of the number. Ways to open up such private-data sources (in a manner that addresses potential privacy and competition concerns) will remain important if we are to be able to hone and improve the analyses. For the moment atleast, consumers of such research have no option but to take the analysis and the results on faith.

8. CONCLUSION

The key advantages of MNBD analysis of human mobility are the higher spatial resolution and the ability to generate frequent forecasts at low cost. In contrast traditional transport forecasts are costly, infrequent and have lower spatial resolution. Traditional survey based techniques do however deliver forecasts that are a comparatively richer source of information in some aspects due to integration with socioeconomic and demographic information. Practically, the way forward will not be to replace these surveys with insights from MNBD, but rather to supplement them. One way this can happen is to calibrate the MNBD based forecasts with insights from the traditional survey instruments at a particular point in time. Then the MNBD data is used to reverse engineer the richer results possible from surveys, for the intermediate times between surveys. The approach retains the strengths of both the traditional and MNBD based forecasting methods by providing frequent forecasts at greater spatial resolution, but which have also been adjusted using the survey data to mitigate many of the issues outlined earlier. Furthermore it may also then be possible for the surveys to be spread further apart in time, resulting in cost reductions.

9. REFERENCES

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

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.

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.

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

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

Telecom Regulatory Commission of Sri Lanka. (2014). Statistical Overview of Telecom Sector. Retrieved September 26th, 2014. Available at http://www.trc.gov.lk/old_site/information/statistics.html.

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